Escuela Politécnica Superior

Departamento de Tecnología Electrónica y de las Comunicaciones

CONTRIBUTIONS TO REGION-BASED IMAGE AND VIDEO ANALYSIS: FEATURE AGGREGATION, BACKGROUND SUBTRACTION AND DESCRIPTION CONSTRAINING

PhD Thesis written by
Marcos Escudero Viñolo
under the supervision of
Prof. Jesús Bescós Cano

Madrid, November 2015

Department: Tecnología Electrónica y de las Comunicaciones
Escuela Politécnica Superior
Universidad Autónoma de Madrid, Spain

PhD Thesis: Contributions to region-based image and video analysis: feature aggregation, background subtraction and description constraining

Author: **Marcos Escudero Viñolo**
Ingeniero de Telecomunicación
(Universidad Autónoma de Madrid)

Supervisor: **Jesús Bescós Cano**
Doctor Ingeniero de Telecomunicación
(Universidad Politécnica de Madrid)
Universidad Autónoma de Madrid , Spain

Year: 2016

Committee: President:

Secretary:

Vocal 1:

The work described in this Thesis was carried out within the Video Processing and Understanding Lab at the Dept. of Tecnología Electrónica y de las Comunicaciones, Escuela Politécnica Superior, Universidad Autónoma de Madrid (from 2009 to 2015). It was partially supported by the Spanish Government trough its FPU grant program and the projects (TEC2007-65400 - SemanticVideo), (TEC2011-25995 Event Video) and (TEC2014-53176-R HAVideo); the European Commission (IST-FP6-027685 - Mesh); the Comunidad de Madrid (S-0505/TIC-0223 - ProMultiDis-CM) and the Spanish Administration Agency CENIT 2007-1007 (VISION).

Part of the work related to the description constraining for point correspondence in wide-baseline scenarios was done while visiting the Center for Intelligence Sensors at the Queen Mary University of London (UK) under the supervision of Prof. Andrea Cavallaro from May 1st to July 31st, 2013.

To my parents, Aida and Javier, who made this possible.

To my sister, Rocío, who has shared, shares, and will share my time.

To Elena, for our past, our present and our future.

Often, in great discovery the most important thing is that a certain question is found.

Max Wertheimer. One of the three founders of Gestalt psychology.

Acknowledgments

First of all, thanks to Pr. Jesús Bescós, for his advices, his help and our long conversations, but, especially, for his support in both the professional and personal areas. Thank you so much Jesús. This Thesis is our work together.

I also want to express my gratitude to Pr. Jose María Martínez for his advice and opinions and for, together with Pr Jesús Bescós, letting me being part of this awesome life experience of teaching and researching at UAM. Thank you so much!

Thanks to Pr. Andrea Cavallaro, his advices and conversations have represented a huge advance in my career. Thanks to Pr. Vasileios Mezaris, Pr. Jenny Benois-Pineau and Pr. Gonzalo Martínez for reading and reviewing this Thesis in its first version and thanks to Pr. Noel O'Connor, Pr. Montse Pardás and Pr. Luis Salgado for being part of the tribunal of this Thesis.

Thanks to all the members in the VPU-Lab, the former (Luis Herranz, Victor Valdés, Fernando López, Victor Fernández, Luis Caro, Jose Antonio Pajuelo and Javier Molina) and the current (Juan Carlos San Miguel, Miguel Ángel Garcia, Fulgencio Navarro, Rafael Martín and Diego Ortego) and specially to Álvaro García Martín, the first person I knew when I started my degree (and still rocking!) and to Luis for believing in my research. Without your support, conversations (and patience) this Thesis will not have been possible. Thank you all! Thanks to all the students from whom I learned how to teach, but specially to Alfonso Colmenarejo, Fulgencio Navarro, Alejandro Blanco, María Narvaéz and Alejandro López; being your *tutor* was one of the best things during these years. Thank you all!

Thank you to all the members in the EPS, professors, P.A.S, people in the cafeteria and to the cleaning personal for creating such a good environment. I feel gratitude for all of them. Thanks to Paulo Villegas, Saúl Labajo and Juan Córcoles for our work together. Thanks to Jorge Ruíz for his advice and conversations and to Pablo Castell, Daniel Ramos and Bazil Taha for their nice words. Regarding the P.A.S. I want to specially say thank you to Marisa, Amelia, Maria José and to all the people in Administration for their efficiency and for turning the bureaucracy in (almost) a pleasant experience. Finally, I want to thank Agustina for her smiles and good work and to José, Diego and Lorenzo for their efficiency (and specially for knowing that I like cold milk in my coffe). Thank you all!

Thanks to my sister, for being always there, for believing in me much more than myself, for her unconditional love and for teaching me so so so many things. Thank you MdD. Rocío Escudero Viñolo. I am so proud of you and so happy to call you sister. Thank you Ernesto, Mamé, María Josefa and Lucnioi, I could not have been done this without your faith in me. Thank you so much abuelos. Thanks to all my family either Viñolo, Escudero, Capel and Delgado. Thanks to my cousins Paula, Álvaro, Rodrigo and Gonzalo and my aunts Raquel, Vanesa, and Salomé and to my uncles Chache, Eduardo and Fernando and of course thanks to my brothers in law (and friends) Pedro, Manu (you changed my life) and Boris, my sisters in law María (thank you for everything) and María Jesús. Thank you Pepa and Pepe, thank you Rosa and thank you Peter, Dani and Sara, for being part of my family. Thank you Antonio, Anlaug and Adela, now I will have time to visit your paradise. Let me specially thank five little persons that immediately produce a smile in my face: Julia, Guille, Louis, Lucía and Candela. Thank you all!

Thanks to my pillars, my friends, for their joy and support. Thanks to Gonzalo, Carlos, Manu, Alex and Rubén, you have been always there and I can not imagine my life without you. Thanks to Marisol, Helena, Nerea, Karol and Mari Carmen I am so happy of having met such wonderful women. Thanks to Belén and Loty for all that have done for me. Thanks to Lurditas, Alfredo, Jo, Julio, Aurelinchi, Laurita, Bombón, Juancho, la Bellas, Ruti, Glauquiño, Andrew, Borjilla, Coque, Miles, Jara, Zuma, Emilio and so many other awesome people and personalities that have severely improved me. Thanks to Marta, Capel, Pascual, Anusa and Campelo, for being always there. Thank you Irene, thank you Diego, you are incredible friends. Thank you Caitlin, you are awesome, and thank you Cotu, for everything, you are all along this Thesis. Thank you all! Thanks to the black liquorice, to Sweden, to Ada and Manuela, to David Simmons and to C.B. Estudiantes, for bringing joy and passion. And if you are reading these lines and believe that your contribution to this Thesis should be mentioned before the black liquorice, you are right, thank you so much for everything that you have done for me.

I especially want to thank the creators of this thesis, those that have provided the tools, the spirit, the support, the ideas and the knowledge that it contains. You both could not have done better. Words are not enough to express my gratitude. Thank you so much to my parents Aida and Javier. I love you.

Elena, you are a coauthor of this Thesis, you have suffered and enjoyed it as much as myself. You are such an incredible woman that your solely name floods my heart of life. *O movimento é vida* and with you, I can not stop moving. I love you; our past is inspiring, our present warm and tender and our future will be full of great news. Thank you for sharing my life.

And last but not least to you, who are reading this Thesis, my most sincere and deep gratitude.

Marcos Escudero Viñolo (November 2015)

Abstract

The use of regions for image and video analysis has been traditionally motivated by their ability to diminish the number of processed units and hence, the number of required decisions. However, as we explore in this thesis, this is just one of the potential advantages that regions may provide. When dealing with regions, two description spaces may be differentiated: the decision space, on which regions are shaped—region segmentation—, and the feature space, on which regions are used for analysis—region-based applications—. These two spaces are highly related. The solutions taken on the decision space severely affect their performance in the feature space. Accordingly, in this thesis we propose contributions on both spaces. Regarding the contributions to region segmentation, these are two-fold. Firstly, we give a twist to a classical region segmentation technique, the Mean-Shift, by exploring new solutions to automatically set the spectral kernel bandwidth. Secondly, we propose a method to describe the micro-texture of a pixel neighbourhood by using an easily customisable filter-bank methodology—which is based on the discrete cosine transform (DCT)—. The rest of the thesis is devoted to describe region-based approaches to several highly topical issues in computer vision; two broad tasks are explored: background subtraction (BS) and local descriptors (LD). Concerning BS, regions are here used as complementary cues to refine pixel-based BS algorithms: by providing robust to illumination cues and by storing the background dynamics in a region-driven background modelling. Relating to LD, the region is here used to reshape the description area usually fixed for local descriptors. Region-masked versions of classical two-dimensional and three-dimensional local descriptions are designed. So-built descriptions are proposed for the task of object identification, under a novel neural-oriented strategy. Furthermore, a local description scheme based on a fuzzy use of the region membership is derived. This characterisation scheme has been geometrically adapted to account for projective deformations, providing a suitable tool for finding corresponding points in wide-baseline scenarios. Experiments have been conducted for every contribution, discussing the potential benefits and the limitations of the proposed schemes. In overall, obtained results suggest that the region—conditioned by successful aggregation processes—is a reliable and useful tool to extrapolate pixel-level results, diminish semantic noise, isolate significant object cues and constrain local descriptions. The methods and approaches described along this thesis present alternative or complementary solutions to pixel-based image processing.

Resumen

El uso de regiones para el análisis de imágenes y secuencias de video ha estado tradicionalmente motivado por su utilidad para disminuir el número de unidades de análisis y, por ende, el número de decisiones. En esta tesis evidenciamos que esta es sólo una de las muchas ventajas adheridas a la utilización de regiones. En el procesamiento por regiones deben distinguirse dos espacios de análisis: el espacio de decisión, en donde se construyen las regiones, y el espacio de características, donde se utilizan. Ambos espacios están altamente relacionados. Las soluciones diseñadas para la construcción de regiones en el espacio de decisión definen su utilidad en el espacio de análisis. Por este motivo, a lo largo de esta tesis estudiamos ambos espacios. En particular, proponemos dos contribuciones en la etapa de construcción de regiones. En la primera, revisitamos una técnica clásica, Mean-Shift, e introducimos un esquema para la selección automática del ancho de banda que permite estimar localmente la densidad de una determinada característica. En la segunda, utilizamos la transformada discreta del coseno para describir la variabilidad local en el entorno de un píxel. En el resto de la tesis exploramos soluciones en el espacio de características, en otras palabras, proponemos aplicaciones que se apoyan en la región para realizar el procesamiento. Dichas aplicaciones se centran en dos ramas candentes en el ámbito de la visión por computador: la segregación del frente por substracción del fondo y la descripción local de los puntos de una imagen. En la rama substracción de fondo, utilizamos las regiones como unidades de apoyo a los algoritmos basados exclusivamente en el análisis a nivel de píxel. En particular, mejoramos la robustez de estos algoritmos a los cambios locales de iluminación y al dinamismo del fondo. Para esta última técnica definimos un modelo de fondo completamente basado en regiones. Las contribuciones asociadas a la rama de descripción local están centradas en el uso de la región para definir, automáticamente, entornos de descripción alrededor de los puntos. En las aproximaciones existentes, estos entornos de descripción suelen ser de tamaño y forma fija. Como resultado de este procedimiento se establece el diseño de versiones enmascaradas de descriptores bidimensionales y tridimensionales. En el algoritmo desarrollado, organizamos los descriptores así diseñados en una estructura neuronal y los utilizamos para la identificación automática de objetos. Por otro lado, proponemos un esquema de descripción mediante asociación difusa de píxeles a regiones. Este entorno de descripción es transformado geométricamente para adaptarse a potenciales deformaciones proyectivas en entornos estéreo

donde las cámaras están ampliamente separadas. Cada una de las aproximaciones desarrolladas se evalúa y discute, remarcando las ventajas e inconvenientes asociadas a su utilización. En general, los resultados obtenidos sugieren que la región, asumiendo que ha sido construida de manera exitosa, es una herramienta fiable y de utilidad para: extrapolar resultados a nivel de pixel, reducir el ruido semántico, aislar las características significativas de los objetos y restringir la descripción local de estas características. Los métodos y enfoques descritos a lo largo de esta tesis establecen soluciones alternativas o complementarias al análisis a nivel de píxel.

Contents

List of Figures

List of Tables

Part I

Part I. Introduction

Contents

This part has two primary targets: introduce the thesis organisation and content and motivate the region for image and video analysis.

To these aims, in chapter 1 we first motivate a region-based analysis by three factors: its semantic potential, its statistical benefits and its relation with human perception. Then, we describe the objective of the thesis and present its major contributions. The chapter ends with a description of the thesis organisation and of its temporal evolution. Chapter 2 is devoted to describe the region on a generic basis. It starts with a definition of the region and of its characterisation. Next, we further motivate the region by presenting its potential use to face several challenges in computer vision applications. The chapter ends by relating these challenges with the rest of the thesis.

"The primitives of a representation are the most elementary units of shape information available in the representation."

David C. Marr and Herbert K. Nishihara. (Representation and recognition of the spatial organization of three-dimensional shapes, 1978)

Chapter 1

Introduction

1.1 Motivation

The semantic potential of regions.

Traditionally, image and video analysis has been conducted at pixel level. The pixel is the smallest analysis unit available in digital visual content; hence, using the pixels as indivisible items to feed the analysis processes seemed a natural solution. This kind of analysis is frequently known as one that relies on punctual operators. Nowadays, the use of punctual operators is disregarded for many applications, as the pixel representation capability is agreed to be strongly constrained. However, we can still find recent methods fully or partially driven by punctual operators, mainly due to computational reasons; for instance, the top-performing background subtraction methods. Nonetheless, the results obtained by these methods are commonly post-processed in a later stage of analysis. In such stage pixel-level results are coalesced to conform spatially-compacted areas on the image lattice. To this aim, the post-processing schemes are usually driven by pixel-adjacency rules which use the spatial arrangement of the image pixels and inter-pixel similarity to shape these areas. Examples of these processes are morphological post-processing (the former) and conditional random field refinement (the latter). Aggregated pixel areas can be considered as regions built in the results domain.

There are several image cues that cannot be obtained at pixel level, but by analysing a pixel neighbourhood; this is usually known as a local-operator driven scheme. For instance, object tracking and people detection are classical applications on which the analysis of isolated pixels is senseless, as both, objects and people, are defined by groups of pixels. Usually, these approximations study rectangular areas of the image. The rectangle—as a generalisation of the square—constitutes a natural unit for local image analysis, as image lattices are generally rectangular-shaped and hence—if the rectangle dimensions are selected in consonance with the image resolution—a complete partition of the image in non-empty sets can be easily achieved.

However, the results of a so-designed scheme sometimes suffer a poor spatial resolution. Results are commonly given on a rectangle-basis, with the rectangle—commonly known as bounding-box—enclosing the pixels which *belong* to the detected object but also several non-object pixels.

On a different scope, holistic analysis of images allows to categorise the images as members of a common space. This kind of analysis relies on global operators and is commonly performed for image indexing and context detection applications with relative success. The image is here classified as a whole, i.e. as a representation of a full scene, without providing any information of the objects that compose it nor of the spatial and functional relations amongst the objects in the image. A scene is composed of objects which are captured at an instant on which they particularly interact amongst them—animated—or coexist—inanimate—. If we aim to derive details of the scene content, we need a local, not a global, description of it.

According to these reflections, the use of regions is here motivated by three premises.

- Post-processing is common in pixel-based methods, and it results in the construction of regions in the results space. We aim to explore the benefits of an inverse operation path, first constructing the regions and then analysing.
- Natural objects are rarely squared nor rectangular. We aim to define image partitions that adapt to the real object contours.
- Scenes are composed of objects which are perceived by grouping pixels information. We aim to divide the objects into intermediate analysis units which are in general smaller than objects but bigger than pixels. Our aim is to check if such division helps in the object characterisation stage.

The statistical benefits of regions.

Consider the following naive example. We are analysing a CGA-frame, which is composed of $320x200$ pixels. We rely on a pre-trained classifier to detect which pixels fulfil a particular condition; for instance, we aim to design an algorithm to detect which pixels are representations —*images*—of scene points which lie on the surface of an specific, previously-trained object.

Let us assume that the designed classifier is a very simple one and just relies on the pixel luminance value to either classify the pixel as an *image* of the trained object or as an *image* of another object—i.e. as a background pixel—. Let us also consider that, in spite of its simplicity, when used in a controlled environment the classifier operates with an accuracy rate of 0.99,—i.e. the classifier returns a correct classification, either object or background, ninety-nine times out of a hundred—.

As the classification of each pixel is an independent process—only per-pixel luminance information is used—, the processes of classifying each frame pixel are stochastically independent events. In these circumstances, the joint probability of correctly classifying all the frame pixels

at the same time is obtained as the product of the individual probabilities. This process leads to a final probability of $0.99^{320x200} \cong 1.3 \cdot 10^{-262}$, i.e. the correct classification of all the image pixels is an extremely rare event in spite of the high accuracy of the individual classifiers.

The luminance can be usually modelled as a discrete variable defined in the range $\mathbf{x} \in [0 : 255] \subset \mathbb{Z}$, i.e. there are only 256 possible values for the luminance. Let us repeat the analysis but by assigning the same classification result to all the pixels that are assigned the same luminance value. In this case, the probability of correctly classifying all the frame pixels—as the individual processes are also stochastically independent events—is $0.99^{256} = 0.0763$. This probability is still quite low, but it is around $6 \cdot 10^{260}$ times higher than the one obtained with the pixel-based classification. Furthermore, note that the gain gets higher as the image resolution is increased.

The classifier and the source of information is the same for both processes, the advantage of the latter respect to the former is only due to the grouping of pixels, i.e. to the building of regions in the luminance space. In this particular example, taking fewer decisions entails making fewer mistakes. Despite the apparent impact of the example, it should be discussed carefully.

First, one can hardly imagine a classifier that, by operating only on punctual luminance information, is characterised by such a high accuracy rate. This is mainly due to the constrained representation capability of the luminance. Although this apparently emphasizes the example conclusions, in fact disregards also a grouping scheme just driven by luminance information, i.e. without considering spatial constraints.

Second, in image and video analysis, the classification processes are rarely independent events, as they search for representations of objects that are continuous in the space and hence, are projected contiguously—but to occlusions—in the image. In fact, this is a clear motivation for spatially driven post-processing methods.

Third, when pixels are grouped, a particular classification process is performed. This process is also defined by an accuracy rate, hence, the aggregated classification error of the grouping stage to the process is rarely zero, as in the example.

Fourth, the error distribution is a factor as relevant as the error accuracy to evaluate a method's performance. In particular, the sparsity or compactness of the errors determines the utility of the obtained results. An interesting discussion about these problems is done in Margolin et al. [2014]. Note that, when using regions, errors are—for better and for worse—usually grouped into spatially compact areas. Consequently, whereas in the pixel-based classification a classification with—to say—10% of erroneous results would still convey useful results, in the region-based classification, and depending on the luminance distribution in the image, the effect would range from no perturbation at all—respect to a 100% successful test—to completely useless results.

In any case, this joint-probability idea motivates us to evaluate the potential improvement achieved when using regions instead or combined with pixel-based classification under two as-

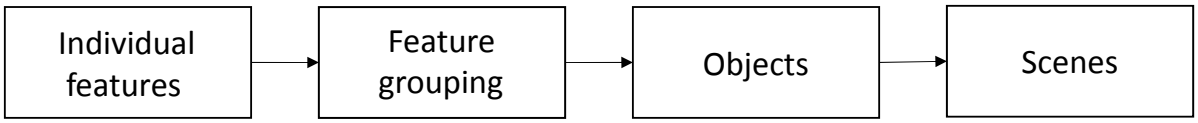


Fig. 1.1. Human perception can be studied at the psychophysical level trough scales: from individual features that conform objects to a whole scene. Adapted from Goldstein [2002]

sumptions:

(1) Pixel grouping into regions is accurate enough.

(2) The decision space on which regions are obtained is *aligned* with the feature space on which classification is performed. Understanding an *aligned* space as one on which the classification problem is *feasible*, i.e. as one on which the features provide an space on which the classes are separable.

Regions in human perception.

Human perception of objects has been studied from several perspectives. Let us focus in the psychophysical perspective (Goldstein [2002]). Under a psychophysical scheme (see Figure 1.1), perception starts by the grouping of individual elementary features—e.g. lines, colours and orientations—. There are several theories that aim to explain how the individual features are grouped and how the grouped features are used to perceive specific objects. Once the objects are perceived, our knowledge of the world or *expertise* is sometimes assumed to drive the rest of the scene perception process.

In our opinion, computer vision is highly inspired by human perception. Computer vision researchers generally aim to provide automatic solutions that derive in comparable scene descriptions to those obtained by objective perception—the subjective sensations associated to perception are out of the scope of this thesis—. This is not necessarily done by replicating or *mimicking* human processing but by proposing methodologies that, starting from similar stimuli—captured light signals—, are able to return semantic descriptions close to human language.

In the following paragraphs, we—warily—aim to motivate the use of regions by enumerating their connections with three of these theories. We agree that different interpretations of these theories are plausible. Nevertheless, these connections have strongly motivated the ideas presented in this thesis and—consequently—their inclusion in this section is necessary to motivate some of the choices made along the document. We here skip the references and detailed descriptions of these theories for the shake of readability; both can be found in chapter 8.

At the end of the XIX century, Wilhemm Wund developed a psychological theory for human perception which is known as structuralism. The structuralists believe that perceptions were

created by the combination of individual elements which they called *sensations*. The clearest example of structuralism processing is probably the *Pointillism*, a technique of painting developed by Georges Seurat and Paul Signac, who were contemporaries of Wund. In *Pointillism*, colour points are arranged on specific patterns to bias the human visual system to automatically *blend* the points and perceive an object. The motivation for region analysis is linked with this theory by relating points or *sensations* with pixels, and regions with the *blending* or combination stage.

The Gestalt theory started from an ambiguity of the structuralism at the beginning of the XX century. In particular, Gestalt psychologists arose to a question: how would the perception of apparent motion—as in an stroboscope—be explained by the sum of individual *sensations*? By contradiction, the well-known statement of the Gestalt theory emerges: *the whole is other than the sum of its parts*. The Gestalt principles—which can be understood as corollaries of this statement—are in fact heuristics for perception, which being not valid for all situations, are *rules* that usually explain the best prediction. Some of these principles are highly connected with the use of regions. For instance, the principle of similarity states that stimuli within an assortment of stimuli are perceptually grouped together if these are similar to each other. This idea constitutes the basis of region-segmentation, which groups pixels that are similar. In the early nineties, additional principles that aim to complete the Gestalt theory were proposed. The principle of the *common region* and the one of the *elements connection* are strong motivations for region-based analysis. The former states that stimuli that are grouped in the same spatial region are perceived together, whereas the latter backs the idea that linked stimuli are perceived as a unit.

At the end of the XX century was proposed the computational theory of David Marr; a theoretical framework to understand object perception. Marr identified three stages in the vision process:

1. A primal sketch: based on the extraction of basic features or fundamental components—mainly edges and groups of edges—.

2. A $2\frac{1}{2}-$sketch: where the fundamental components are combined to shape scene surfaces and texture is acknowledged.

3. A 3D$-$sketch: on which objects are shaped as composed of basic partially-schematic primitives.

The theory is rich in useful schemes and experimental ideas that have influenced researchers in computer vision up to our days. In particular, our methods are partially motivated by the grouping steps that are required to evolve from one sketch to the following.

On one hand, —again strongly linked with the Gestalt principles—the theory states that the simple features that conform the primal sketch are grouped to shape the $2\frac{1}{2}-$sketch. This grouping includes several sub-stages, including the study of image contours that limit the object

surfaces. A similar motivation arises when studying the feature-integration theory of attention developed by Treisman.

On the other hand, the last sketch, the object centred 3D−sketch which is assumed to fulfil perception, is claimed to be composed of basic partially-schematic primitives. These primitives are claimed to ease the recognition and may or may not constitute full-meaning entities. The recognition-by-components theory of Biederman is also highly linked with this premise.

1.2 Objectives

The main objective of this thesis is to explore the benefits and limitations of using regions in some stages of image and video analysis. We try to leverage the use of regions as a potential alternative or complement to pixel-based analysis. To achieve this objective we propose to study the region following a three stages philosophy, with each stage associated to specific objectives:

- Region definition.
  - We aim to provide a proper definition of the region both in terms of its semantic content, its description capabilities and its potential use.
- Region construction.
  - We aim to explore and organise existing approaches to perform region construction, commonly known as region-segmentation, so that our contributions can be contextualised.
  - We aim to enhance mode-seeking for region-segmentation via setting strong, but plausible, constraints derived from applying the scale-space theory.
  - We aim to provide flexible tools for local variability description through filter-bank basis-functions. To this aim, we define processes for the automatic selection of basis-functions and the generic comparison of filter responses.
- Region-based applications.
  - We aim to explore and organise existing approaches for background subtraction so that our contributions can be contextualised.
  - We aim to improve pixel-based background subtraction methods by exploring region-based schemes.
  - We aim to derive a part-based object identification methodology by combining regions with local descriptions to fight occlusions.
  - We aim to improve the matching of image points in wide-baseline scenarios by defining projective transformations of fuzzy-characterisation schemes.

1.3 Major contributions

The significant novel contributions of this thesis are summarised as follows:

1. An alternative definition of region, according to:
   (a) name origin.
   (b) basic characterisation.
   (c) potential applications.
2. An updated dual organisation of region-segmentation approaches:
   (a) analysis stages: pre-processing, feature extraction, local analysis, globalisation and regionalization.
   (b) level of processing: local, global and combined processing.
3. An automatic spectral-bandwidth selection scheme for Mean-shift region-segmentation of luminance images, that allows:
   (a) a scale-space analysis for mode detection.
   (b) a variable bandwidth according to global distribution.
   (c) a faster convergence of the local seeking processes.
4. A perceptual metric to measure the similarity between responses of the Discrete Cosine Transform filter-bank, which is composed of:
   (a) an study of the representativeness of the filter responses in the description of natural images.
   (b) a metric to compare any two sets of filter responses.
5. A robust to illumination region-segmentation scheme, based on Mean-shift, to cope with shadows and lit areas in background subtraction, which operates by:
   (a) inserting reflectance consistency in the Mean-shift mode-seeking stage.
   (b) searching for colour alignment in the Mean-shift mode fusion stage.
6. A region-driven scheme to face and model background dynamism by covariance-based updating, which relies on:
   (a) a framework to combine multiple description features based on inter-feature correlation.

(b) a multi-layer scheme to account for both temporal variations and region-segmentation inconsistencies.

7. A part-based object identification approach which, requiring very low training, is robust to severe object occlusions, and is conformed by:

   (a) an study of the connections of existing object identification methods with human perception theories.

   (b) a single layer to organise knowledge based on the inter-similarities of part-based descriptions.

   (c) a scheme to define self-adaptive description areas for state-of-the-art local descriptors.

8. A flexible and adaptable method to match image points in wide-baseline scenarios, based on the consideration of projective transformations of the local descriptor area, which uses:

   (a) a self-adaptable support definition for local characterisation by studying the sparsity of inter-pixel similarities.

   (b) an automatic constraining of the possible projective deformations of the description support.

   (c) an affinity-weighted feature transformation of the description support.

Additional minor contributions of this thesis are summarised below:

1. A review and a novel organisation of background subtraction methods.

2. A feasibility study of the use of the Discrete Cosine Transform for the modelling of the temporal evolution of background and foreground samples.

3. A multi-layer and multi-class technique for background subtraction.

1.4 How to read this document

Parts organisation and contents

This document is structured in five parts and appendixes, which are organised as follows:

- Part I: Introduction.

  – *Chapter 1: Introduction.* This chapter presents the motivation, the objectives, the main contributions and the structure of the thesis.

  - *Chapter 2: Regions in image and video analysis.* This chapter introduces the region, motivates its use and presents some of its potential applications.

- Part II: Regions as feature aggregators.

  - *Chapter* 3: *Region segmentation.* This chapter contextualises the designed methods for aggregating pixels into regions by proposing an organisation of existing region-segmentation approaches.
  - *Chapter* 4: *Mean-Shift Region segmentation based on the automatic bandwidth selection in the scale-space.* This chapter describes a methodology to automatically select a bandwidth for every input sample in any technique based on Mean-Shift.
  - *Chapter* 5: *Local variability modelling via the Discrete Cosine Transform.* This chapter proposes a novel metric on the DCT filter-bank to measure the local-variability around a pixel.

- Part III: Regions for background subtraction.

  - *Chapter* 6: *Challenges, key-tasks and recent trends in background subtraction.* This chapter contextualises the proposed contributions respect to the state-of-the-art in background subtraction approaches.
  - *Chapter* 7: *Contributions to region-driven background subtraction.* This chapter describes the proposed region-based contributions to background subtraction: robustness to local illumination and handling of background dynamism.

- Part IV: Regions for description constraining.

  - *Chapter* 8: *Severe-occluded object identification via region-based descriptions.* This chapter proposes a region-driven constraining approach of two-dimensional and three-dimensional local descriptions and exemplifies its use for the task of object identification in severe-occluded scenarios.
  - *Chapter* 9: *Projective deformation and appearance transformation of region-supports for wide-baseline point matching.* This chapter proposes a point matching scheme for wide-baseline calibrated scenarios which relies on the generation of a subset of projective-deformations of a region-support around each anchor point.

- Part V: Conclusions and future work.

  - *Chapter* 11: *Achievements, conclusions and future work.* This chapter concludes this document summarizing and concluding the main results obtained and motivating future research lines.

- VI: Appendixes. Here we include relevant research results achieved while exploring region-based approaches, but not directly related to regions.

  - *Appendix A*: *A feasibility study of the use of the DCT for the task of background subtraction.* This appendix motivates the hypothetical use of the DCT-based metric proposed in chapter 5 for background modelling and foreground detection in background subtraction approaches.
  - *Appendix B*: *Multi-class background subtraction.* This appendix presents a pixel-based background subtraction algorithm that relies on a multi-layer and multi-class background modelling to store the temporal evolution of dynamic backgrounds.
  - *Appendix C*: *Super-pixel based isolation of the Scale invariant feature transform* . This appendix presents a solution to constrain the SIFT local description by using super-pixels.

A representative diagram of the contents of the thesis is depicted in Figure 1.2.

Suggested reading order according to the reader profile.

The thesis has been structured in parts that are organised according to a two-level design: region-segmentation and region-based applications. Furthermore, contributions in region-based applications are further divided into two parts: background subtraction and local description. However, as each part is quite independent of each other, the thesis admits alternative reading orders. Nevertheless, we suggest the reader to start with Part I—as you are doing as otherwise you would haven't reached this suggestion—and to finish with Part V in order to end with an overall conclusion of the thesis.

For instance, a reader solely interested in region-segmentation may find some of the ideas presented in Part II agreeable and can skip all the other parts. If the reader searches for contributions to background subtraction, we recommend a fast reading of chapter 6 and then the rest of Part III. Then, the reader can consult the two first appendixes—maybe after a quick review of the beginning of chapter 5. Finally, if the reader is concerned with local description we would suggest to start with the comparison stage of chapter 5, then continue with the simple constraining scheme of appendix *C* and next with the whole Part IV. These three examples of reading orders are depicted in Figure 1.3.

Temporal evolution of this thesis.

The evolution of this thesis has not being linear, as it frequently happens in when exploring new scientific paths. Whereas the motivation for the use of regions has been always the core of the thesis, it started with an study of the state-of-the-art in background subtraction. We then

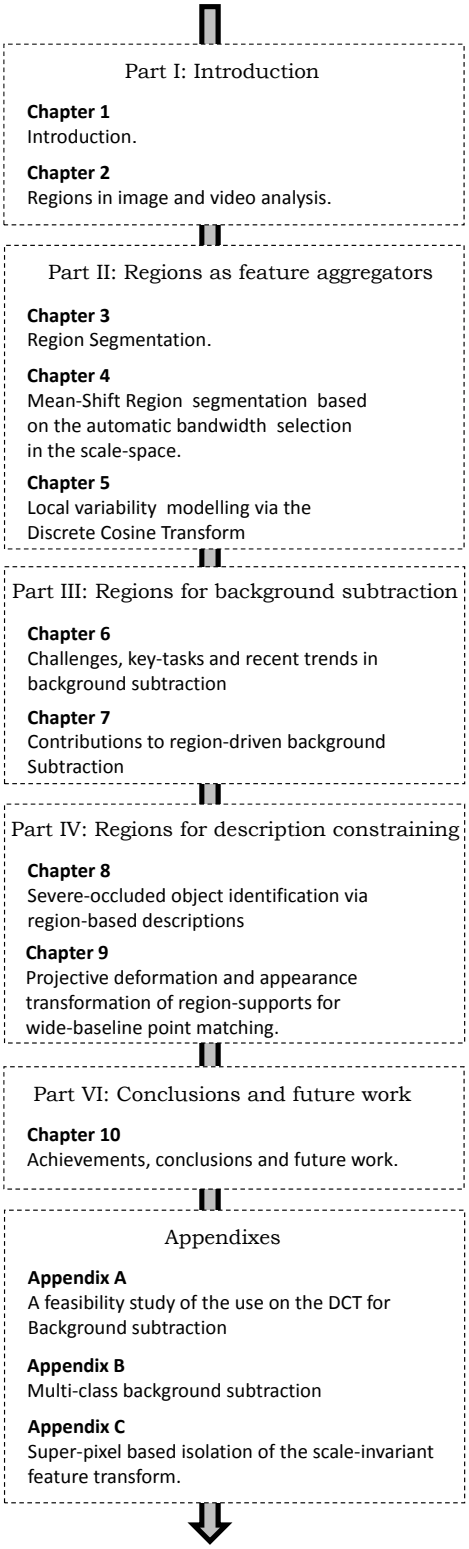


Fig. 1.2. Diagram of the contents of the thesis.

studied the existing approaches for region-segmentation. These two studies led to the combined solutions proposed in chapter 7. The problems that we found motivated us to improve their operation. We faced these problems by two different strategies. One was devoted to improve background subtraction, firstly by dealing with camouflage through local-variability modelling, resulting in the solution described in Appendix *A* and, secondly, by inspecting alternative background modelling schemes, leading to the algorithm described in Appendix *B*. The other strategy was devoted to improve region-segmentation, firstly by diminishing the number of parameters of region segmentation methods, which led to the solution described in chapter 4, and then by studying local-variability modelling (which is explained in chapter 5). In parallel, and aside for research in areas not related with this thesis, we studied local-characterisation constraining. It all started with a segregation idea. If we aim to describe individual points according to their surroundings, and the surrounding of these points partially change; the characterisation would no longer be valid. Under this idea, we developed the SP-SIFT description described in Appendix *C*. Evolutions of this idea, but focused on specific applications are proposed in chapters 8 and 9. The temporal evolution of the thesis is sketched in Figure 1.4.

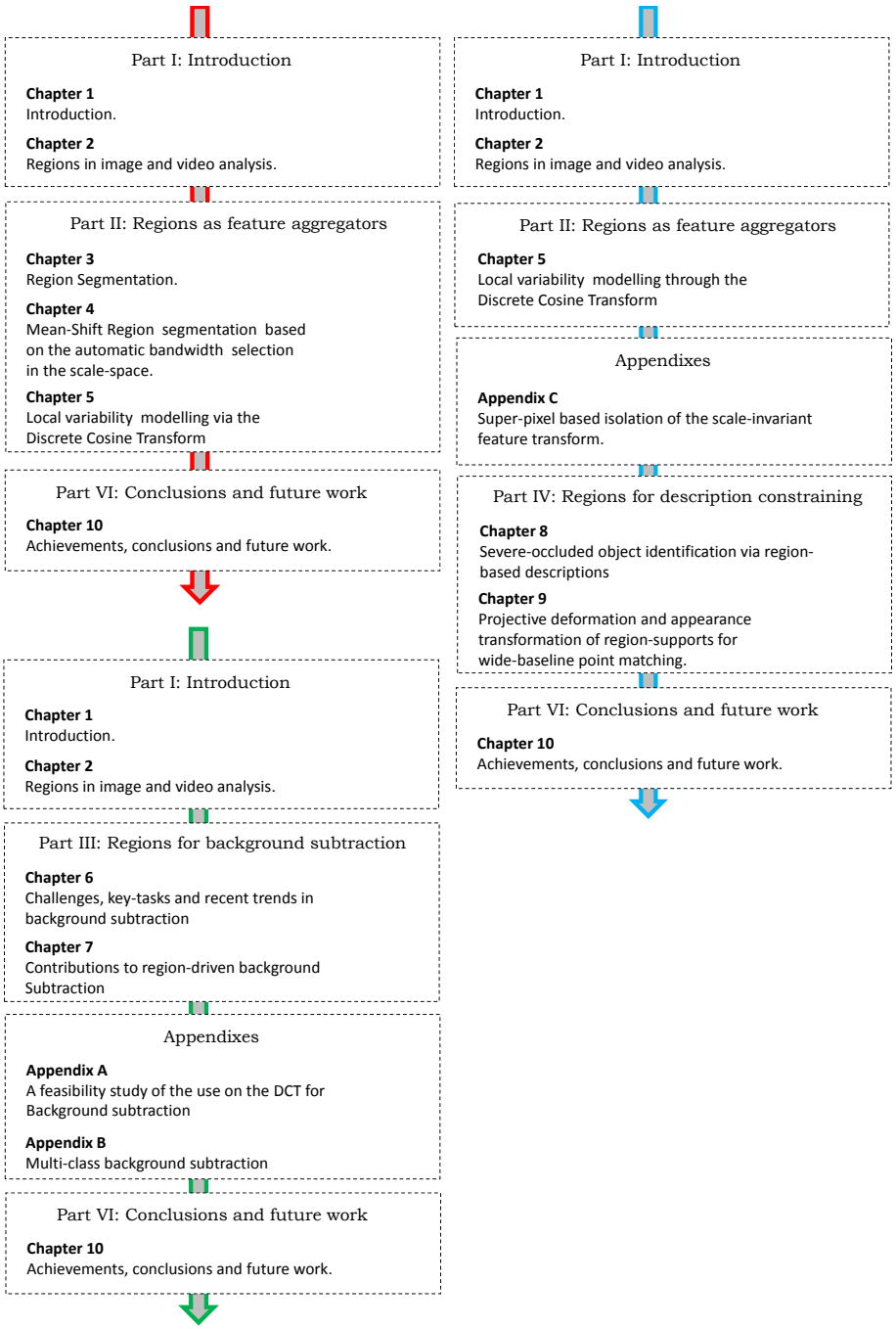


Fig. 1.3. Suggested reading order according to the reader profile. A reader interested in region-segmentation (top left, indicated by a red arrow), background subtraction (bottom left, indicated by a green arrow) or local description (right, indicated by a blue arrow) may be specially interested in some of the chapters of this thesis.

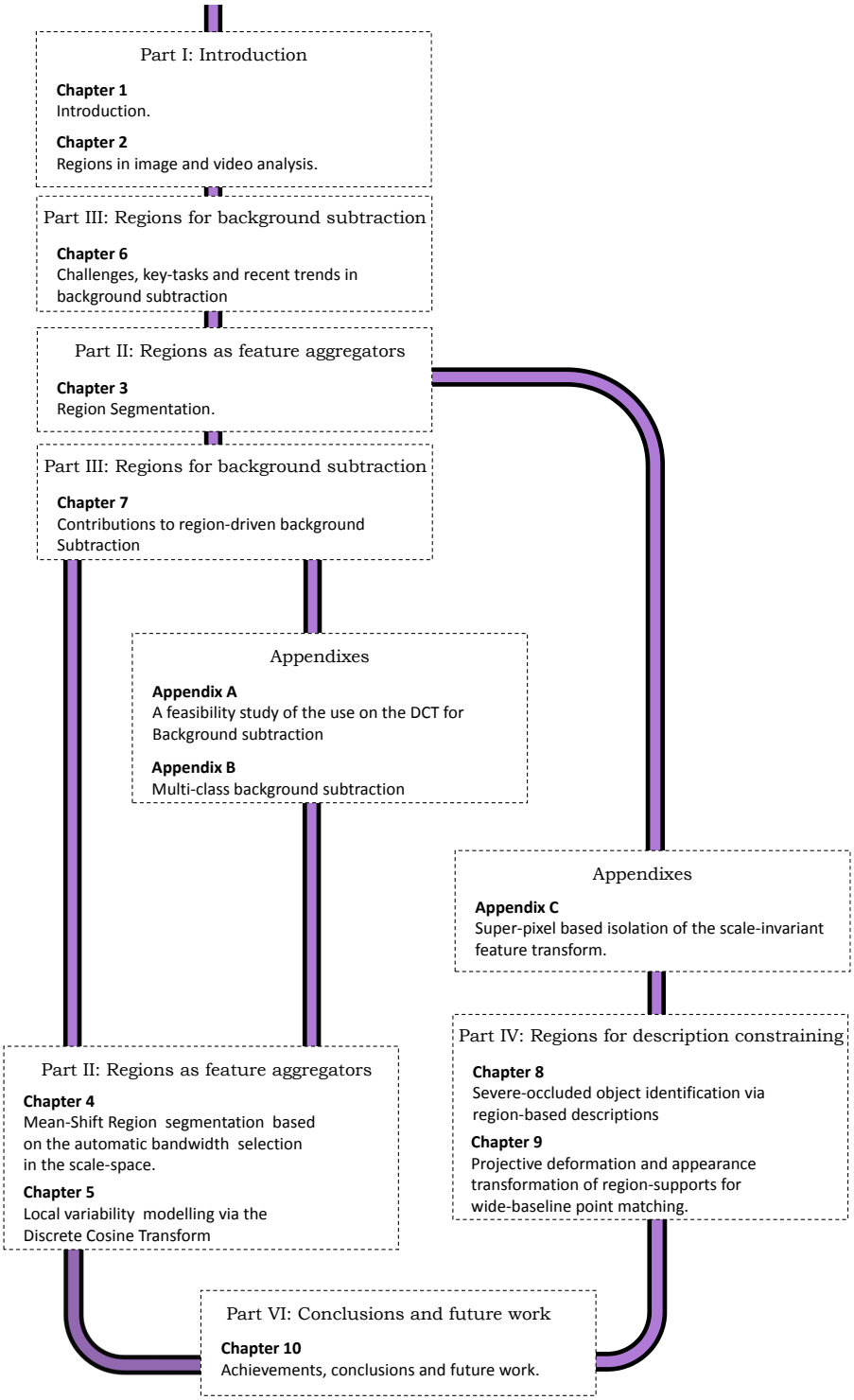


Fig. 1.4. Temporal evolution of the thesis. Time starts on top and advances downwards.

Chapter 2

Regions in image and video analysis

In this chapter we aim to motivate the use of the region for image and video analysis. To this aim, we first introduce the region conceptually and explore the associate characteristics that automatically result from the region definition and can be used to represent it. Then, we review some of the analysis challenges that motivated its emergence and definition. Finally, we introduce the uses of the region explored along this work.

2.1 Semantic and practical definitions of a region.

Definition of the region

In spite of the huge amount of studies devoted to define, use and organise region-based image and/or video analysis approaches—a subset of them are reviewed in chapter 3—it is hard to find a generic definition of the region concept. Before analysing region segmentation approaches and exploring the potential applications of the region, we think that the reader would benefit from a preliminary definition.

Let us start discussing this question by brief reviewing (four of) the definitions of the word *region* that we can find in widely used online English dictionaries:

- *region*: any large, indefinite, and continuous part of a surface or space [1].
- *region*: an area considered as a unit for geographical, functional, social, or cultural reasons[2].
- *region*: an open connected set together with none, some, or all of the points on its boundary [3].

[1] *first definition in the Online Collins dictionary*: http://www.collinsdictionary.com/
[2] *second definition in the Online Collins dictionary*: http://www.collinsdictionary.com/
[3] *sixth definition in the Online Merriam-Webster dictionary*: http://www.merriam-webster.com/

- *region*: range, area, or scope [4].

We can observe that these definitions are quite varied among them, but somehow, together, they coalesce to a common meaning.

region: any large, indefinite, and continuous part of a surface or space.

The first definition is the most specific amongst the four included. The definition implies that a region should be: *large* and *continuous* on one side but also *indefinite* at the same time. Aside for this—arguably—contradiction, the definition identifies a region as a part of a surface or a space. This is a very interesting implication—for our discussion—, as it locates the region in a (even) larger whole. Therefore, as a corollary for this fact, we can understand that the whole (the space) might contain several regions or, in other words, that this whole can be divided into regions.

region: an area considered as a unit for geographical, functional, social, or cultural reasons.

The second definition may seem the less related with image and video analysis. However, let us restate the definition excluding some adjectives: *an area considered as a unit for functional reasons*. Understanding functional as: *designed to have a practical use* [5] we have achieved a perfect motivation—from an engineering point of view—for the region. Furthermore, the region is here defined as a unit: *any group or individual, esp when regarded as a basic element of a larger whole* [6], thus reinforcing the intuitions derived from the first definition.

region: an open connected set together with none, some, or all of the points on its boundary.

The third definition is the key one. Under a topology-wise definition, the Merriam Webster dictionary defines a region as: *an open connected set...,* with *set* being *a group of objects with stated characteristics*[7]. Furthermore, the nature of the set is also specified: a connected set. The definition also explain the type of objects that compose the set latter on: *....together with none, some, or all of the points....* Then, the region is a group of connected points—assuming that all the objects in the regions are of the same nature—. This fact, together with the previous definitions, places the region as an intermediate grouping level between points and space, that is, up to this point, a region can be defined as it follows.

region: a group of connected points that represents a part of a space.

[4] *fifth definition in the Online Collins dictionary*: http://www.collinsdictionary.com/
[5] *first definition in the Online Merriam-Webster dictionary*: http://www.merriam-webster.com/
[6] *second definition in the Online Collins dictionary*: http://www.collinsdictionary.com/
[7] *Online Cambridge dictionary* http://dictionary.cambridge.org

Moreover, the definition is closed with a very important statement: *...on its boundary.* Therefore, the region is associated with a boundary—it is an open set if it does not contain the boundary or a close set otherwise—. Whereas a boundary is *a point or limit that indicates where two things become different* [8]. Walking on eggshells to avoid formal fallacies, we can deduce that the points in the region are less different amongst them than to the points that are not in the region. Therefore, the points in the region are more *similar* amongst them that to the connected points that are not in the region.

region: range, area, or scope.

The fourth and last definition is the most generic and is prone to be subjected to intensive corrections by a rigorous mathematician. Nevertheless, it is the first to include the term range: *the extent included, covered, or used* [9], then inherently partitioning the space in covered and uncovered parts. Therefore, if the space is a whole to which the region represents a part, the rest of the space can be defined as the complementary of this part. With this in mind, we can derive a more specific, albeit still generic, definition of region:

region: a group of *similar* connected points that represents a part of a space which is *dissimilar* to a region in its complementary part.

This definition is of course subjected to how similarity (or dissimilarity) is measured and to the nature and dimensionality of the space. However, it constitutes a keystone on which building the rest of this chapter.

Definition of region segmentation

If the whole space is to be divided into regions and we impose that no part of the space can be leaved unassigned, we are performing a partition of the space in disjoints sets (regions) whose union is the space itself. This process is commonly known as region segmentation, where the word segment is used as a generic term inherited from common geometrical terms, e.g. line segment, sphere segment or cylindrical segment.

region segmentation: a partition of the space in disjoints regions whose union is the space itself.

As it has been defined, the region segmentation process can be easily expressed in mathematical terms:

[8] *third definition in the Online Merriam-Webster dictionary*: http://www.merriam-webster.com/
[9] *sixth definition in the Online Merriam-Webster dictionary*: http://www.merriam-webster.com/

Being Ω the space and $\mathcal{P}_n(\Omega)$ a partition of the space in n parts: $\Omega_j, j = 1, ..., n$, $\mathcal{P}_n(\Omega)$ is a region segmentation process if:

$$\mathcal{P}_n(\Omega) = \left\{ (\Omega_1, ..., \Omega_n) \ : \ \Omega = \bigcup_{j=1}^{n} \Omega_j \text{ and } \Omega_j \cap \Omega_k = \emptyset \text{ for all } j \neq k \right\} \tag{2.1}$$

, with each part, Ω_j , being a region.

The space Ω on which the region segmentation process $\mathcal{P}_n(\Omega)$ operates is known as the *decision space* (Salembier and Marques [1999]). This space is composed of d-dimensional points $\mathbf{x}$, with d being the dimensionality of the *decision space*.

Region segmentation and clustering

Up to this point, the obtained definition for both the region and the region segmentation process are almost equivalent to those usually used to define a cluster and a clustering process (e.g. see the definitions in Jain et al. [1999]). So, what are the differences between a region and a cluster? In general terms, and if the segmentation is performed in an unsupervised manner, there is just an essential difference between them: the spatial arrangement of the points. Regions are usually extracted on image and video content, where points are arranged on spatial and temporal lattices. Some region segmentation approaches disregards this fact—we categorize these as clustering approaches (see chapter 3) as may not hold with our previous definition of a region in its *connection* requirement—. However, points spatial arrangement is a key cue for object perception (Goldstein [2002]); hence, it should be considered when partitioning the space.

Representations of the region

A region can be represented in several ways, the simplest being the label j that identifies the region in the partition (see equation 2.1). However, additional characterizations might provide more information about the region and about the points grouped in the region. Some of these characterizations are usually available—or can be extracted with little effort—after the segmentation process.

The region label

All the points grouped in a region Ω_j by a segmentation process $\mathcal{P}_n(\Omega)$ would share a common identifier j , with j being the region label.

That, is, being $lb_\Omega(\mathbf{x}) : \mathbb{R}^d \rightarrow \mathbb{R}^1$ a labelling function:

$$lb_\Omega(\mathbf{x}) = j \text{ for all } \mathbf{x} \in \Omega_j \tag{2.2}$$

The label image

Let $\mathbf{I}$ be an image organised in a 2−dimensional array in the Cartesian discrete plane $\mathbb{Z}^2$. Let the *decision space* Ω be a set of feature points $\mathbf{x}$— e.g. the image luminance values or the image colour vectors—each of them describing each pixel $\mathbf{p} = (u, v)$ in $\mathbf{I}$.

We can define $\mathbf{Lb_\Omega}$ as a label image of the same size of $\mathbf{I}$ which contains at each pixel coordinates (u, v) the label assigned by a region segmentation process $\mathcal{P}_n(\Omega)$ to the point on that position in $\mathbf{I}$: $lb_\Omega(\mathbf{x})$.

Even if spatial constraints (e.g. the (u, v) coordinates of each pixel) are included in *the decision space* Ω, several unconnected pixels in $\mathbf{Lb_\Omega}$ might be assigned the same label due to partition errors. In order to achieve a partition in agreement with the achieved definition for region—as well as for practical and semantic reasons—it is usually preferred to operate with connected-component regions, that is, with regions that constitute individual areas in the image domain. This can be achieved by a connectivity analysis on the label image (Haralock and Shapiro [1991]).

The connected-component version of the label image

Let us first define the concept of connected-component.

Being $\mathbf{Mk}_{\Omega,j}$ a binary mask of the same size of $\mathbf{I}$ resulting from activating—setting to 1—all the pixels assigned to label j in $\mathbf{Lb_\Omega}$ and de-activating—setting to 0—the rest of the pixels, a connected component is a set of activated connected pixels in $\mathbf{Mk}_{\Omega,j}$.

Two pixels $\mathbf{p}_0$ and $\mathbf{p}_s$ in $\mathbf{Mk}_{\Omega,j}$ are connected if it exists a path of pixels $(\mathbf{p}_0, ..., \mathbf{p}_i, ..., \mathbf{p}_s)$ such that for all $0 \leq i \leq s$, $\mathbf{Mk}_{\Omega,j}(\mathbf{p}_i) = 1$ and $\mathbf{p}_{i-1}$ and $\mathbf{p}_i$ are neighbours in the image plane, i.e. in $\mathbb{Z}^2$. Hence, to define connectivity we also need to define the concept of neighbouring.

Let $\mathcal{N}_4(\mathbf{p})$ be the 4−neighbourhood in $\mathbb{Z}^2$ of a given pixel $\mathbf{p} = (u, v), u, v \in \mathbb{Z}$:

$$\mathcal{N}_4(\mathbf{p}) = \{(u+1, v), (u-1, v), (u, v+1), (u, v-1)\} \tag{2.3}$$

, and $\mathcal{N}_8(\mathbf{x})$ its 8−neighbourhood in $\mathbb{Z}^2$:

$$\mathcal{N}_8(\mathbf{p}) = \mathcal{N}_4(\mathbf{p}) \cup \{(u+1, v+1), (u-1, v-1), (u-1, v+1), (u+1, v-1)\} \tag{2.4}$$

Two pixels $\mathbf{p}_{i-1}$ and $\mathbf{p}_i$ are neighbours in $\mathbf{Mk}_j$ under an 8−connectivity premise if $\mathbf{p}_i \epsilon \mathcal{N}_8(\mathbf{p}_{i-1})$. Along the rest document we use 8−connectivity as the default connectivity.

Repeating this process for every mask $\mathbf{Mk}_j, j = 1, ..., n$ and assigning the same label to the pixels in the same connected component and different labels to pixels in different connected components, a relabelled image $\mathbf{Lb_{\Omega,\mathbb{Z}^2}}$ is obtained. This label image combines the results of the region segmentation process with the connectivity constraints imposed on the image domain. Furthermore, this label image implicitly defines a new partition: $\mathcal{P}_{n^{cc}}(\Omega)$:

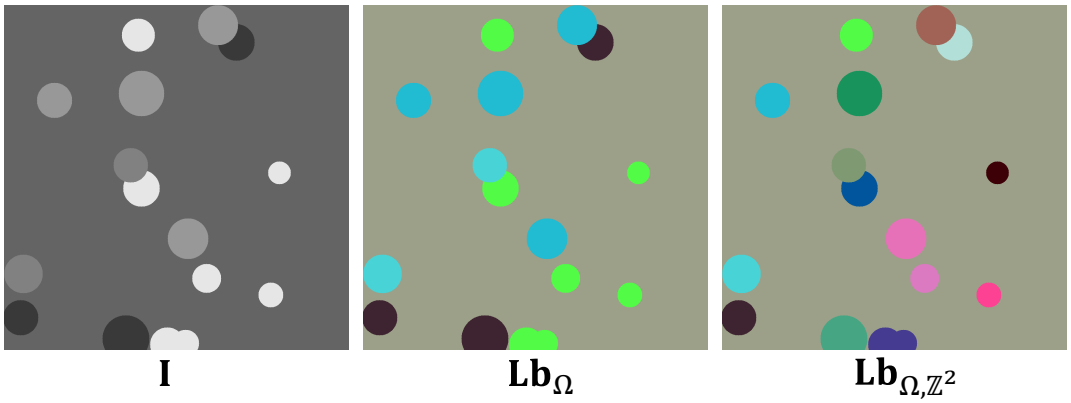


Fig. 2.1. Labelling region segmentations. Luminance image $\mathbf{I}$ (left) is segmented in regions and its label image $\mathbf{Lb_\Omega}$ (centre) is obtained. A connected component analysis is performed on $\mathbf{Lb_\Omega}$ to obtain $\mathbf{Lb_{\Omega,\mathbb{Z}^2}}$ (right). Note how not connected regions in $\mathbf{Lb_\Omega}$ are assigned new labels in $\mathbf{Lb_{\Omega,\mathbb{Z}^2}}$. Labels have been here represented by random colours to ease visualization.

$$\mathcal{P}_{n^{cc}}(\Omega) = \left\{ (\Omega_1^{cc}, ..., \Omega_{n^{cc}}^{cc}) \ : \ \Omega = \bigcup_{j=1}^{n^{cc}} \Omega_j^{cc} \text{ and } \Omega_j^{cc} \cap \Omega_k^{cc} = \emptyset \text{ for all } j \neq k \right\} \tag{2.5}$$

, with each part, Ω_j^{cc} , being a connected-component region and with $n^{cc} \geq n$.

Note that this process is exportable to higher-dimensional Cartesian spaces $\mathbb{Z}^N$—as in videos, where $N = 3$ with time as the third dimension—by first defining a proper neighbourhood for each pixel.

An example of a label image $\mathbf{Lb_\Omega}$ obtained without spatial constraints and of its associated connected-component version $\mathbf{Lb_{\Omega,\mathbb{Z}^2}}$ are included for comparison in Figure 2.1 .

The region representative

The similarity between two points $\mathbf{x}$ and $\mathbf{y}$ in the *decision space* Ω can be measured through a pair-wise function $D_\Omega(\mathbf{x}, \mathbf{y})$ which is also commonly associated to Ω.

As aforementioned, every point $\mathbf{x}$ grouped in a region Ω_j must fulfil a similarity criterion:

$$D_\Omega(\mathbf{x}, \mathbf{y}) \leqq \varepsilon, \text{ for all } \mathbf{y} \in \Omega_j \tag{2.6}$$

, and a dissimilarity criterion at the same time:

$$D_\Omega(\mathbf{x}, \mathbf{z}) > \varepsilon \ , \mathbf{z} \in \Omega_k \text{ for all } k \neq j \tag{2.7}$$

, where we have assumed that D_Ω is a metric—which is usually the case—and then Ω is part of a metric space: $\Omega^* = (\Omega, D_\Omega)$.

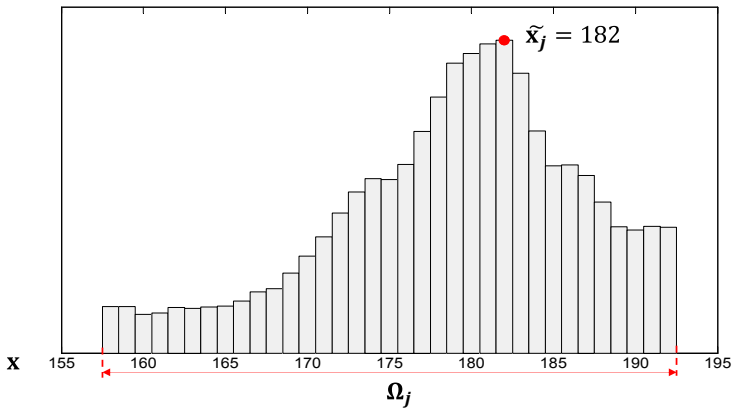


Fig. 2.2. Example of region mode estimation $\tilde{\mathbf{x}}_j$ inside a region Ω_j for *1*-dimensional points.

If $\varepsilon = 0$, the region is homogeneous in the *decision space*, thus, any point in the region would identically represent the region, i.e. every point can be the region representative, as all the points in the region are in fact *the same* point in the *decision space*—respect to D_Ω metric—.

If, otherwise, the region is not entirely homogeneous, $\varepsilon > 0$, a proper representative ($\tilde{\mathbf{x}}_j$) of the region can be obtained by several methods, including the arithmetic mean of all the points in the region:

$$\tilde{\mathbf{x}}_j = \frac{1}{|\Omega_j|} \sum_{\mathbf{x} \in \Omega_j} \mathbf{x} \tag{2.8}$$

, where $|\Omega_j|$ is the cardinality of the region Ω_j, or the median of the points in the region:

$$\tilde{\mathbf{x}}_j = \begin{cases} \mathbf{y}_{(|\Omega_j|+1)/2} & , |\Omega_j| \text{ , } odd \\ \mathbf{y}_{(|\Omega_j|)/2} + \mathbf{y}_{(|\Omega_j|+1)/2} & , |\Omega_j| \text{ , } even \end{cases} \tag{2.9}$$

, with $\mathbf{y}_{(|\Omega_j|+1)/2}$ and $\mathbf{y}_{(|\Omega_j|)/2}$ being the $\left(\frac{|\Omega_j|+1}{2}\right)$ and $\left(\frac{|\Omega_j|}{2}\right)$ ordered statistics of the points in Ω_j.

Alternatively, the representative can be also extracted as the most common point in the region, i.e. as its mode. In order to extract the mode, the discrete empirical distribution of the points in Ω_j should be first computed. This is usually achieved through an histogram on the range of the points in Ω_j.

Figure 2.2 exemplifies this process for 1−dimensional points on an hypothetical region which contains points in the range $\mathbf{x} \in [158, 192]$. This process leads to a mode: $\tilde{\mathbf{x}}_j = 182$. Note that, for comparison, the arithmetic mean value for the same region is $\tilde{\mathbf{x}}_j = 178.9$, whereas the median value on the same region is $\tilde{\mathbf{x}}_j = 180$. Note also that, among the three schemes for extraction of the region representative discussed, the mode is the only one that ensures that the

Fig. 2.3. Image (left) and its region RGB-mode representation (right). Original image is part of the LabelMe data-set (Russell et al. [2008]).

representative equals a feature point, i.e. $\tilde{\mathbf{x}}_j \in \Omega_j$, with independence of the region cardinality.

Figure 2.3 includes an example of region segmentation of a colour image with each region being described by its mode. In this case, modes were extracted in the RGB colour space (Kuehni [2003]). In the region image, colour transitions represent region transitions.

An interesting corollary of the proposed region definition, is that it should be a unique mode on the region points empirical distribution—i.e. the distribution should be uni-modal—, as otherwise the points in the region wouldn't be *similar* among them. This fact is used for mode-seeking approaches to perform the segmentation $\mathcal{P}_n(\Omega)$, as we review in chapter 3 of this document.

The region boundary

The extraction of the boundary of Ω_j in the *decision space* does not provide, in general, information about the region spatial extent on the image support. Instead, it is usually preferred to extract the region boundaries in the image domain. To this aim, we can operate on the label image $\mathbf{Lb}_{\mathbf{\Omega},\mathbb{Z}^\mathbf{2}}$ and use the neighbourhood definition in equation 2.4.

The boundary of a region Ω_j is the set of pixels E_j in $\mathbf{Lb}_{\mathbf{\Omega},\mathbb{Z}^\mathbf{2}}$ such that for every pixel $\mathbf{p}$ in the set, its neighbourhood $\mathcal{N}_8(\mathbf{p})$, contains at least one pixel labelled as j and one pixel not labelled as j.

Note that, if all the pixels in the set E_j are labelled as j, the boundary is part of the region. Conversely, if none of the pixels in the set E_j is labelled as j, the boundary is external to the region (in this case, the boundary circumscribes the region). Two regions, j and k, that share a boundary $E_{j,k}$ are considered adjacent regions. The graph that arranges all the regions in a partition regarding their adjacency relationships is known as a region adjacency graph (RAG).

Fig. 2.4. Extraction of the region boundary. First row: an RGB image (left) is segmented in regions leading to a set of labels which are here represented by random colours for visualization(right) . Second row: the RGB colour mode (right) and the boundary (left) can be extracted on the region. The boundary is here shown over-imposed on the RGB image.

An example of the extraction of the region boundary is included in Figure 2.4. Note that, as the region boundary has been defined, it confines the region—either internal or externally—. This fact is used for contour detection approaches to bypass the segmentation process $\mathcal{P}_n(\Omega)$, as we review in chapter 3 of this document.

2.2 Motivation for region segmentation approaches

The motivation for performing region segmentation is hard to be discussed in generic terms as it irremediably converges to practical issues. Several studies and surveys on the topic have explored the hypothetical benefits of performing region segmentation Salembier and Marques [1999]; Cheng et al. [2001]; Freixenet et al. [2002]; Zhang et al. [2008a]; Kim and Hong [2009]; Ilea and Whelan [2011]; Vantaram and Saber [2012]; Abdelsamea et al. [2014]. The following discussion partially emerges from the ideas there described, but it is also the result of our own reflections.

In general, region segmentation is predominantly employed as a preprocessing step either to annotate, enhance, (pre) analyse, classify, categorize, and/or abstract information from images. However, in the scope of image and video analysis, we have identified five challenges that motivate the use of regions. Let us discuss each challenge separately.

Narrowing the semantic gap

Under a somewhat crude abstraction, one of the main objectives of image and video understanding techniques is to overcome *the semantic gap* . The *semantic gap* is defined by Smeulders et al. [2000] as the empty space between the human interpretation of content and the representation of such content as a digital video signal .

In other words, it is the difference between the formulation of contextual knowledge in a powerful language (e.g. the human natural language) and its formulation in a logical formal language (e.g. the pixel-based colour representation in a video).

In practical terms, for example, when interacting with image or video content, people would like to access information (searching, indexing, viewing or tagging it) via high level scene descriptions, e.g. with a description of the image content, instead of with a description about the pixel characteristics and arrangement.

Imagine that *the semantic gap* is pictorially represented vertically, as a cliff, instead of horizontally, as bridging-the-gap approaches do (Zhao and Grosky [2002]). Climbing up the cliff would place us closer to the top of the cliff and at the same time further from its bottom. Under this metaphor, the concept of *semantic gap* can be extended to that of *the semantic pyramid.*

The *semantic pyramid* can be understood as a division of *the semantic gap* into several levels of understanding. For instance, in an image signal, the lowest level in the pyramid would be the pixel level; pixels can be grouped to form the region level; regions can be further grouped to conform the object level and, finally, the scene or group of interrelated objects would constitute the top of the pyramid. Note the similarities of this scheme with the one depicted in Figure 1.1 of chapter 1.

Analysis techniques aiming to generate high level semantic descriptions should ascend in the pyramid starting from the lowest level, i.e. the one available in the first term: the pixel-level. Moving on to the region level would constitute a narrowing of *the semantic gap*, as we are supposed to be one step closer to the human language. This hypothetical narrowing may be somehow supported by human perception theories. A review of these theories, in the scope of object perception, can be found in chapter 8.

In any case, the practical benefits associated to this narrowing theory are twofold. First, regions boundaries usually coincide with object boundaries. A region partition might oversegment objects which are not flat in the space analysed, but the limits that separate these objects from their background are usually conserved by the segmentation process. Second, a

Fig. 2.5. Object decomposition into regions. The *Glico Man* (and the sun at his back) is here decomposed into regions. Regions RGB modes are included in the top row whereas region labels (with random-generated colour codes) appear in the bottom row. A region segmentation in a *decision space* Ω composed only of the pixels colour vectors would group pixels according to their colour characteristics. This hypothetical grouping is here represented by the column-wise organisation of the regions. However, it is often preferred to group pixels so that these conform connected areas in the image domain, which is the approach here used as represented by the region labels.

region partition of an object represents a description of such object in basic shapes (see Figure 2.5). Together, these two properties provide an abstraction of the image space which is: usually more reliable—fitted to the objects contours—as the similarity criteria are more restrictive and, usually simpler—composed of a lower number of regions— as these criteria are relaxed.

An optimal selection of these criteria hypothetically conveys the simpler image abstraction which conserves the object contours. This abstraction is not only potentially easier to process, but might also contain relevant information about the structure of the objects.

Signal de-noising

In some of the graphical examples provided up to this point (Figures 2.3 and 2.5) we have represented each region by its representative (a colour vector). Under this representation, the semantic meaning of the region—the human interpretation of the area enclosed by the region—remains unaltered.

By means of this process, each data point has been represented as function of its nearby points in the *decision space*. This constitutes a local smoothing process which has the intrinsic ability of partial data de-noising. The majority of the noise included in the process of image

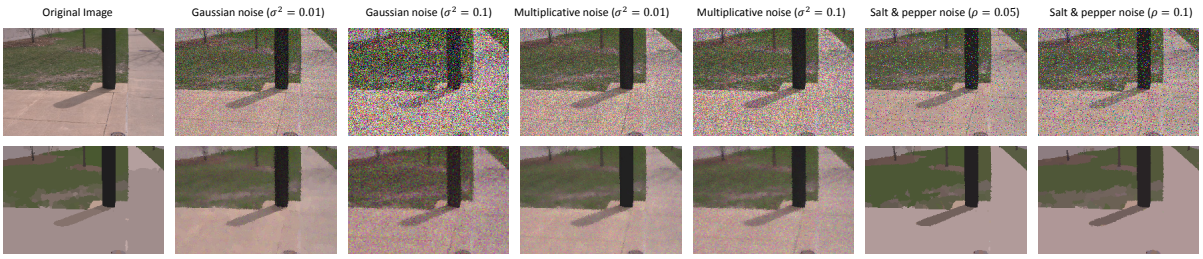


Fig. 2.6. Image de-nosing ability of region segmentation. Top row. RGB colour images affected by several kinds of noise. Bottom row. Region segmentation results of the EDISON Mean-Shift based approach with spectral and spatial bandwidth equal to 10 (see chapter 4 for details). Original image is part of the shadow-oriented data-set in Guo et al. [2013].

capturing is usually assumed to be confined in the high level frequencies of an image. Such noise share frequency ranges with the fine detail of the image. Under a noise-removal basis, region segmentation approaches can also be viewed as de-noising schemes that aim to conserve the image fine detail. The success in this process will be defined by the grouping criteria and the noise intensity and distribution. However, in general, region segmentation entails data de-noising. This ability is subjected to the condition that the captured noise does not severely affect the signal structure—i.e. to the condition that the signal-to-noise ratio is high enough so that the process result is determined by the signal data, not by the noise—.

Figure 2.6 includes an example of this idea. In the example, we have added artificial noise of different types and with different *distributions* to a given image. We then use the EDISON system [10] to perform region segmentation. The EDISON system operates on colour images and combines Mean-Shift (Comaniciu and Meer [2002]) and edge detection (Meer and Georgescu [2001]) to perform synergistic region segmentation (Christoudias et al. [2002]).

Results in Figure 2.6 suggest that in situations on which the noise does not occlude the image information, a region-segmentation process is able to diminish or even completely eliminate its influence. However, when the noise is strong enough to occlude image information, the process is unable to solve it.

Space sampling

Let assume that some kind of analysis operation (classification, detection, etc.) is aimed to be performed on points in a *feature space*—which may or may not overlap with the *decision space*—. Let also assume that a preliminary region segmentation is available, and remember that a region Ω_j has been defined as a group of points that share some common properties.

We aim to compare the problem complexity from two perspectives: pixel-driven and region-driven. Under these premises, let us formulate two questions:

[10] http://coewww.rutgers.edu/riul/research/code/EDISON/

(a) Should all the points in a region Ω_j be assigned equal results of the test?

And, in the case of an affirmative answer,

(b) Is the analysis of one of the points in Ω_j , or better, of the representative of Ω_j , enough to attain the same result for all the points in the region?

From the discussion up to this point, we can state that an affirmative answer to question (a) implies that the analysis result itself is a common property of the points in the region Ω_j in the *feature space*. Or. in other words, that the *feature space* is correlated (or aligned) with the *decision space*. Hence, if (a) is answered affirmative (b) should also be so.

The implications of this double affirmative answer are of high relevance. As the region is supposed to represent a cleaner version—up to some degree—of its points, the influence of noise in the operation results is diminished. Due to this noise reduction effect, the distribution of the analysis results would be more compact than a direct analysis on the points. Finally, the number of operations to be performed can be drastically decreased.

The decreasing rate depends on the grouping criteria and the *decision space* structure. For instance, in the example included in Figure 2.3, the number of regions obtained by the segmentation and connected component analysis is 8068, compared with the $480 \times 640 = 307200$ pixels in the original image: the number of regions represents less than the 2.7% of the original number of pixels.

In practical applications, regions are obtained on a *decision space* and under similarity criteria so that they are supposed to provide affirmative answers to the question (a). This inherently constitutes a sampling of the space into a subset of points—usually the region representatives—each one representing one of the regions in the partition.

Adaptable description supports

How is the neighbourhood of an image pixel? Unlike the colour or the brightness, some features, as texture, need to be defined on a spatial support. Regions appear as an alternative to fixed supports on which to measure the distribution of spatial-based features. The assumption commonly made is that a region is prone to define a common entity at a—generally unknown—semantic level. This entity can be used as the support for extracting alternative features that may be useful to take later decisions at higher semantic levels of analysis.

The benefits of using regions as description supports—compared to using fixed supports—are mainly summarized in a description-independence premise: as regions are prone to conserve object contours, the influence of contiguous objects in the description is eliminated. This might be specially useful when describing moving objects in videos. There, the object moves in a scene, while the objects around it remain static or move differently to it. Moreover, it might be also beneficial when, with independence of the object dynamics, the scene is captured from

two different points of view. In this case, as the scene is 3−dimensional and the image plane is 2−dimensional, the captured configuration of the scene may vary substantially; changing completely the adjacency and occlusion relationships among the captured objects.

Fuzzy encoding of inter pixel similarities

In the proposed definition of a region (equation 2.5) it is assumed that a point can be a member of just one region. That is, that the point-to-region membership relation is binary (*hard*): a point is either in a region or not in a region.

In a fuzzy segmentation process, each point belongs to several—to all in the limit of the meaning—regions at the same time, with a fuzzy membership degree to each one.

Fuzzy (*soft*) segmentations may retain more image information than *hard* segmentations in some cases. For instance, when analysing digital images, a scene boundary between two (objects) parts may not be correctly captured in the image due to the quantization process performed by digital CCD sensors. Hence, pixels representing this boundary are prone to contain information of both parts, creating new structures that may or may not be continuous in the image domain. In these situations, a fuzzy relationship of membership of these pixels to the regions representing the parts delimited may benefit the classification of these pixels.

Additionally, the similarity criterion ε to create regions may be set different for different areas in the image—for instance, due to the discussed noise distribution or to the concept of scale, on which we deepen later in the document—. In fuzzy segmentations this criterion is leaved out of the process and is applied *a posteriori* in a segmentation stage commonly known as *defuzzification*. This process has been proved to substantially reduce the influence of noise (see Gong et al. [2013]).

In our opinion, fuzzy regions can provide special benefits when using the regions as supports for descriptions. If one aims to describe an individual pixel and to use the region to which it has been assigned as support for description, fuzzy membership can be used to define levels of similarities between the pixels inside the region and the pixel to be described. Through this scheme, the influence of potential errors committed during the segmentation process can be reduced.

2.3 Challenges explored along this thesis.

Chapters 3, 6 are devoted to review and organise existing approaches in the tasks of region segmentation and background subtraction, therefore, they can be considered contextualising chapters.

All of the discussed challenges motivated the designed techniques described in the rest of the chapters up to some degree. However, the region is obtained or used differently in each chapter

motivated by a specific subset of challenges. Let us associate each chapter with the challenges that motivated its development.

Chapter 4 describes a region segmentation process which relies on a pre-detection of the scale to drive the segmentation. The result is an image partition on which the influence of noise has been reduced. The solution in chapter 5 is explicitly designed to conserve the fine detail of the image while further reducing the noise influence (both captured and semantic). Its potential is exemplified by its application to the task of object boundary detection in natural images. Together, these processes are based on the signal de-noising principle of regions and are designed to produce a boundary map whereby reduce the semantic gap.

On a different scope, chapter 7 uses regions to drive a foreground detection process. It starts from the assumption that the temporal evolution of a pixel is a noisy signal severely affected by captured and semantic noise. In the chapter, it is assumed that the region constitutes a robuster analysis unit that the pixel. Moreover the region is also used to extract spatially-sampled descriptions. Therefore, this chapter covers several region principles: signal de-noising ability, space sampling and adaptable supports.

The objective of chapter 8 is to identify objects in cluttered scenarios where scene surface information is available. This is achieved by characterising the objects through region-masked surface normals, thereby using regional supports for description and sampling the scene space. Finally, chapter 9 weaves together the fuzziness and the adaptable support ideas in order to describe points in a wide-baseline scenario.

Part II

Part II. Regions as feature aggregators

Contents

This part addresses region segmentation, i.e. the stage devoted to obtain regions from the image content.

First, in chapter 3, we review relevant approaches in the state-of-the-art of region segmentation. We propose to arrange these approaches according to their features and their operation strategy. Regarding the latter, we divide approaches into global, local and combined depending on how these create regions. We explore the main existing techniques and present a generic flowchart for region segmentation. The chapter ends with a brief discussion on data-sets and evaluation metrics. Chapter 4 is inspired by a future research line proposed in Comaniciu et al. [2001]. There, the authors closed the paper by sketching a unification of the scale-space theory and Mean Shift; up to our knowledge such unification remains unexplored. In the chapter we modestly propose a scheme to integrate both schemes under strong but plausible assumptions. Finally, chapter 5 presents a new method to handle local-variability in natural scenarios in order to detect scene contours. Related state-of-the art methods usually rely on the design of crafty spatial filters. There is lack of research in the formalisation of the number, nature and scale of the spatial filters used for analysis. Furthermore, it is unclear how responses of these filters should be compared. We propose a method that operates on the DCT, a singularity-blind filter-bank—e.g. one that is not aligned with specific edges orientations—and derive a metric to compare the response of any two of the filters in the filter-bank. The potential advantages of the designed method are exemplified by its use for the generation of an image contour map.

"Parts and wholes evolve in consequence of their relationship, and the relationship itself evolves."

Richard C. Lewontin and Richard Levins (The Dialectical Biologist, 1985)

Chapter 3

Region segmentation

Having discussed the potential benefits of a region-based analysis in chapter 2, the vast amount of efforts devoted and algorithms developed to perform region segmentation should not strike us. The organisation, study and proper evaluation of the existing approaches may well constitute a thesis by its own.

There is a considerable number of excellent region segmentation surveys in the literature: Salembier and Marques [1999]; Cheng et al. [2001]; Freixenet et al. [2002]; Zhang et al. [2008a]; Kim and Hong [2009]; Ilea and Whelan [2011]; Vantaram and Saber [2012]; Abdelsamea et al. [2014]. However, due to the fast and extensive evolution of region segmentation, some of these surveys deal about approaches which are rarely used nowadays (Freixenet et al. [2002]; Cheng et al. [2001]; Zhang et al. [2008a]). Furthermore, the scope of some other surveys is too broad (Vantaram and Saber [2012]) or too focused on a particular sub-field of the topic (Salembier and Marques [1999]; Kim and Hong [2009]; Ilea and Whelan [2011]; Abdelsamea et al. [2014]) to yield an exhaustive, yet generic, view of unsupervised region segmentation methods. This situation inhibits the constitution of these studies as representative surveys on the topic.

In this chapter we propose a generic organisation of recent as well as classic unsupervised region segmentation approaches that have stood the test of time. The chapter is intended to provide a flexible classification to allow the inclusion of future methods rather than to constitute an extensive listing of all the existing—most unused—approaches.

The chapter starts with a discussion on relevant topics which are needed to state the overall problem. Then, the proposed organisation is presented and the studied categories are briefly described and exemplified by relevant approaches of the state-of-the-art. Following, a listing of the relevant data-sets and quality measures used to evaluate the goodness of unsupervised region segmentation approaches is presented. Finally, a set of conclusions is derived.

3.1 Prior discussions.

The region segmentation is an ill-defined problem.

In 2001, the Computer Vision Group of the Berkeley University made public an evaluation data-set (Martin et al. [2001]) which was the result of an study about object contour identification by humans. Human observers were asked to break up a given scene into pieces that represent *distinguished* things in the image, with all the pieces having approximately the same relevance. In Figure 3.1, the annotations of five images in the data-set are included for visualization. From the annotations, two conclusions were derived in Martin et al. [2001]:

1. Some images may be uniquely segmented whereas some others accept multiple solutions.

2. The variability of the solutions was mainly due to differences in the *level of attention* from one human observer to another.

On one hand, some of the users provide annotations which were really tight to the real object contours, even when these were not well-defined due to a lack of contrast. However, their previous knowledge of the world and their semantic sapience allowed them to construct the regions even in these situations. These processes are part of the active vision paradigm, e.g. of the ability of the vision system to selectively control the image acquisition process (Aloimonos et al. [1988]). These capabilities are still ahead of artificial vision systems.

On the other hand, whereas users agreed in some regions, specially those defining the foreground (most salient) part of the image—see how two users just annotated the stair in the third image, disregarding the background structure—, some of the annotations did not coincide. Several other factors can be well discussed, including the reasons that led, four of the five users, to fuse the mountains profile in the last row of Figure 3.1. However, for our purposes, a more relevant reflection arises.

If the agreement among users exists, but it is not reiterative, how would a human observer rate a particular segmentation obtained by an automatic segmentation method? And, leaving aside the subjective evaluation, how should the quality of a segmentation be quantitatively evaluated? These questions have been recursively studied in the literature. We discuss some of the most relevant reflections and ideas there taken in section 3.6 of this chapter.

In any case, up to this point, we aim to instil the idea that the region segmentation task is one of a high evaluation complexity, as even human users do not agree about what a region is.

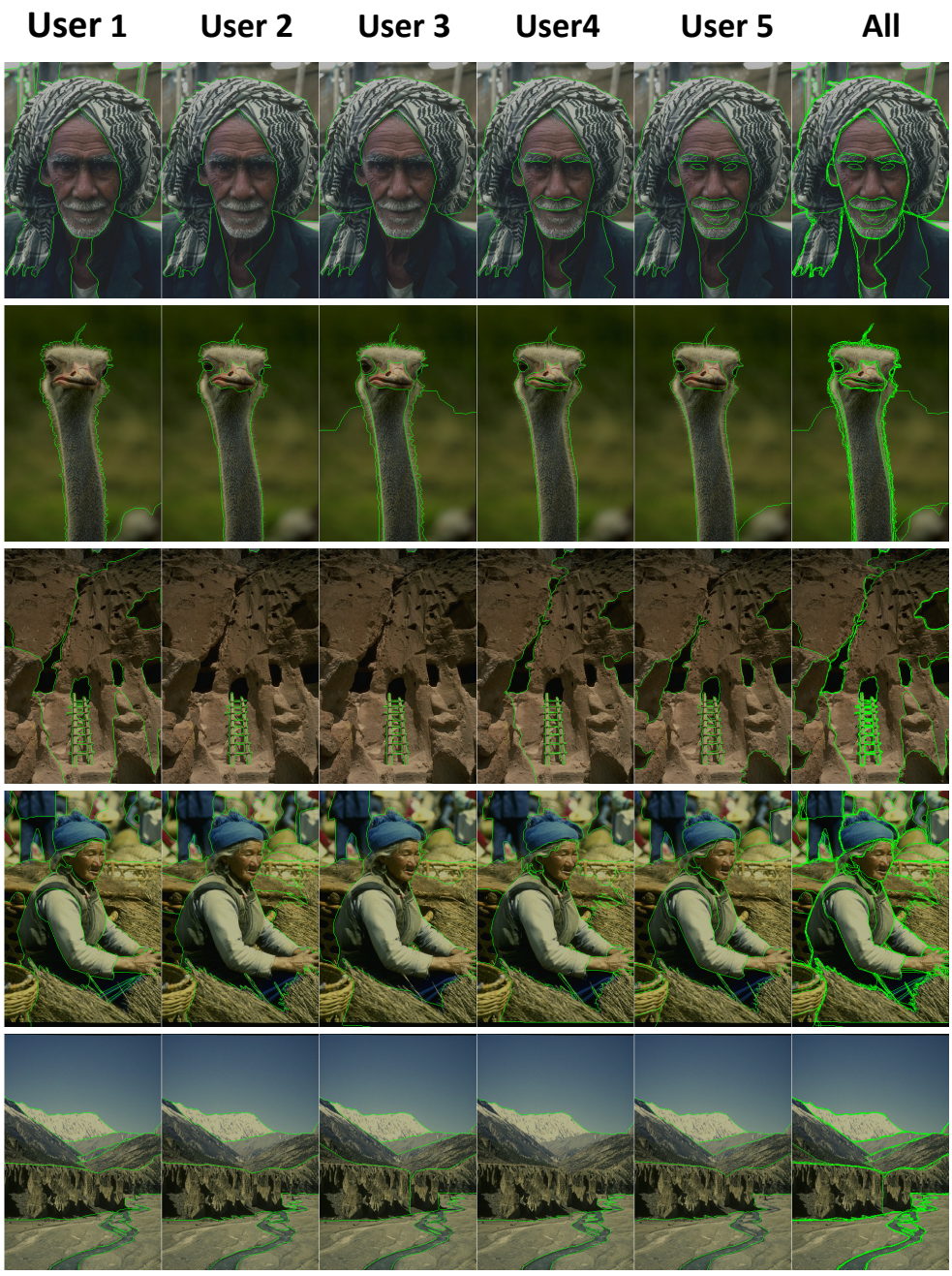


Fig. 3.1. Some examples from the Berkeley data-set (Martin et al. [2001]; Arbelaez et al. [2011]). Region boundaries are superimposed—in green—on each RGB image. First five columns show annotations for five different users. Top right column agglutinates the combined annotation of all the users. Note how different regions are identified in the images according to the semantic interpretation of each user.

Features.

Regions are defined by their homogeneity and by their transitions to neighbouring regions. These transitions can be observed in the luminance, in the colour or in the texture space. However, some of the transitions may be observed in some spaces but not in others. For instance, the turban in the first row of Figure 3.1 is defined by a texture transition. If we instead focus just in colour and luminance transitions the turban might be over-segmented, i.e. divided into regions of a lower semantic interpretation.

Furthermore, the presence of uneven illumination inside a region, the perspective and scale distortions and the influence of image noise are common factors that increase the complexity of finding the relevant transitions and, hence, the regions in an image.

Several spectral features have been used to drive region segmentation approaches, being the favourite the RGB colour (Hoang et al. [2005]; Shi and Funt [2007]; Gong et al. [2013]; Comaniciu et al. [2001]; Felzenszwalb [2004]). However, some authors prefer to convert the original RGB image to alternative colour spaces. The perceptually uniform CIELab is used in Achanta et al. [2012]; Ugarriza et al. [2009] and Arbelaez et al. [2011] among others. The reason behind this use is the belief that operating in a colour space which is adapted to the human perception of colour would benefit the detection of regions. The difference between two CIELab colour vectors is related to the human reactions to colour transitions. Consequently, by operating in the CIELab colour space, a region segmentation approach would be on even grounds with humans, at least, in their interpretation of colour changes. Alternative colour spaces to RGB and CIELab have been rarely used for region segmentation, but there are approximations which also include the YIQ colour space in their analysis space (Ilea and Whelan [2008]). Furthermore, in Mignotte [2008], RGB, CIELab and YIQ are used together with other three colour spaces (HSV, XYZ and LUV) in a simple but effective approximation.

Most of these methods are—theoretically—able to operate also on the luminance space. Alternatively, the scope of some approaches is limited to the analysis of this space as it is supposed to contain all the spectral transitions of an image, which is not necessarily true—as discussed in Barnard et al. [2002]—but often enough to obtain accurate results (Mumford and Shah [1989]; Chan et al. [2001]; Sharon et al. [2006]; Sawatzky et al. [2013]).

As aforementioned, the use of texture descriptions is mandatory to detect some edges. There are several schemes to describe textures, including Gabor filters (Hoang et al. [2005]) and local binary patterns (Randen and Husoy [1999]; Qian et al. [2011]). However, in the context of contour detection, *Textons* have been the preferred option (Martin et al. [2001]; Malik et al. [2001]; Arbelaez et al. [2009, 2011]). In the scope of image analysis, the *Textons* were defined in Malik et al. [2001] as prototype responses to a set of predefined spatial filters. In particular, the *Textons* are there obtained by quantifying and clustering the multidimensional responses to a set of predefined spatial filters during a training stage.

Finally, as discussed in chapter 2, existing region segmentation approaches aim to provide connected component regions as outputs of their processing. This is achieved either by the region segmentation process itself—by incorporating spatial information in the process or naturally provided by the nature of the process—or by post-processing techniques on the resulting segmentation.

Edges, contours, boundaries and regions.

As pointed out by Martin et al. [2004] and Arbelaez et al. [2011], the edges and boundaries extraction techniques are related, but are not identical, mainly due to one reason: edges are only defined at feature level, that is, they do not delimit entities as regions or objects. Consequently, edge detectors do not necessarily return closed contours, also known as boundaries, which, instead are straightly defined by the output of region segmentation approaches.

Summing up, an edge can be defined as a change in the analysed feature, while a boundary is a delimiter between the projections of scene surfaces. A region is the area enclosed by a boundary whereas a contour can be both, an edge or a boundary, the former if it is open, the latter if it represents a closed pixel-path. Nevertheless, if edges are extracted by studying size-adequate neighbourhoods, the region partition of the image is latent in the edge information, albeit incomplete. Regions can thus be obtained by post-processing closing techniques on the edges. Ultimately, we can find two directions in the extraction of the image boundaries: from regions to contours and *vice versa*. Relevant studies have been done in the task of contour detection; let us discuss them briefly on a chronological basis.

Classical edge detectors: Marr and Hildreth [1980]; Roberts [1963]; Duda et al. [1973]; Prewitt [1970]; Canny [1986], detect edges by thresholding the convolution of local derivative filters of different nature over the grey scale or luminance image; therefore these are blind to edges responding only on colour or texture features. In Freeman and Adelson [1991] these approaches are extended by introducing a quadrature pair of even and odd symmetric filters at different scales and orientations, thus incorporating multi-scale analysis in the process.

Anisotropic diffusion also represents a strong field of research in the task of edge detection (Perona and Malik [1990]; Chao and Tsai [2010]). Basically, approaches in this area model diffusion over some image feature—generally, the luminance—as a heat diffusion process: the intensity would spread inside flat areas but not over contours. Their objective is two-fold: noise reduction and edge sharpening. In Lopez-Molina et al. [2014], anisotropic diffusion methods are exhaustively reviewed.

However, all of these schemes are blind to texture edges, as they search for continuity on a spectral feature. In Randen and Husoy [1999]; Paclik et al. [2002]; Drimbarean and Whelan [2001]; Qian et al. [2011] it is shown how including texture descriptions substantially improves the detection of edges. However, as these approaches rely only on texture information, they fail

to detect edges just described by luminance or colour transitions—e.g. edges between flat areas. As both approaches are complementary, it seemed natural to combine them.

The well-known studies described in Martin et al. [2004] and Dollar et al. [2006] jump from edge to contour detection by combining multiple cues by linear regression. Both algorithms rely on the extraction of several cues either hand-crafted (Martin et al. [2004]) or of high simplicity (Dollar et al. [2006]). In Martin et al. [2004], the output of the edge detection filters defined in Malik et al. [2001] is combined with brightness, colour and texture gradients. These gradients are extracted by comparing the histogram-distribution of these features on the two halves of a fixed radius circular area around a pixel. The diameter that generates the halves is rotated several fixed angles in order to respond to differently oriented contours. Authors evaluate several combinations of these cues and select an optimal combination of the parameters by supervised training. In Dollar et al. [2006] the features and their combination in Martin et al. [2004] are seen as manually tuned and designed on purpose for the evaluated data-set. Instead, authors propose to generate a big set—around 50000 features—and train their response to boundaries on several scenarios by Probabilistic Boosting Trees. Results indicate that their method is able to adapt to varied scenarios. Moreover their contour detection responses have a probability associated—differently than in Martin et al. [2004]— which is argued to be in consonance with human perception of edges.

In Ren [2008], a set of contrast-based features similar to those used in Martin et al. [2004] are extracted at multiple scales. The multi-scale extraction substantially outperforms previous methods as allows to identify fine and coarse detail in the image at the same time. We will go back to multi-scale analysis later in this chapter.

Finally, in Leordeanu et al. [2012], a unified formulation for boundary detection is presented, but through a different approximation. The study is motivated by the aim to avoid the dependency of previous methods to oriented filters (Malik et al. [2001]; Martin et al. [2004]; Arbelaez et al. [2009, 2011]). To this aim, they define a method to detect, at the same time, the intensity and the orientation of the contour. The method relies on the efficient creation of carefully designed matrices that combine low and mid-level cues. The contour detection problem is then solved by a single Eigenvalue decomposition.

Bottom-up vs top-down.

Bottom-up and top-down information processing paths can be also used to describe and organise region segmentation approaches.

In the context of region segmentation (Vantaram and Saber [2012]), a top-down approach would start from the whole space and obtain regions by the disaggregation of space parts that do not fulfil desired criteria. This process is commonly known as region splitting. Bottom-up approaches, instead, start form the smallest available units, e.g. the pixels descriptions, and

sequentially merge them by a region merging strategy.

Whereas bottom-up approaches are still receiving substantial attention by the research community, top-down approaches are—nowadays—rarely used in their original spirit. Nonetheless, as an example of a classical well-established top-down processing it is worth to mention the Watershed transform (Roerdink and Meijster [2000]), which is still being used mainly as a post-processing tool: Ilea and Whelan [2008]; Arbelaez et al. [2011].

However, top-down processing still lies in the core of several region segmentation approaches; either in the shape of energy minimisation functions that operate on the whole *decision spac*e, or by complementary holistic techniques that are used to aggregate/disaggregate local decisions on a region hierarchy. In order to distinguish these new trends in top-down processing from classical splitting techniques, we tag them as globalization approaches.

Multi-scale analysis.

Let us refer again to Figure 3.1 and discuss some of the incongruous user annotations. We can see that the incongruities among users are specially associated to regions defining the details of the objects, e.g. the eyes and the eyebrows of the man in the first row. On the other hand, it is interesting to observe that—instead—the facial features of the woman in the fourth row are not annotated by any user—. Neither by those who identify these of the man as relevant parts. This may be well explained by the concept of scale, i.e. the relative size of these features respect to the scene.

In words extracted from Lindeberg [1993]: the details of an image only exist as meaningful entities over limited ranges of scale. In Koenderink [1984], the scale problem is illustrated by a division of the object extent into two scales, the *inner scale* and the *outer scale*. The *outer scale* encapsulating the whole object and the *inner scale* as the set of scales at which different substructures (details) of the object can be observed. Let us illustrate this division with an example: a tree would constitute an object in the *outer scale* whereas its branches may be structures at different scales of the *inner scale*.

The image content is fixed, i.e. we can not go closer or further to an object in an image once it has been captured without altering the original data, and with it, the objects structure. Therefore, when analysing a given image—or a frame in a video—only a certain range of scales are available, i.e. certain structures may not be recoverable after captured. However, if the image is analysed at different scales—i.e. with different spatial extents—a higher number of details, shapes and structures can be recovered.

3.2 Proposed organisation of region segmentation approaches.

The number of region segmentation approaches has increased exponentially since the early nineties (Ilea and Whelan [2011]). A search for *region segmentation* in Google Scholar currently results in almost 1.7 millions of results. For comparison, 230 results were indexed in 2008 by Ilea and Whelan [2011]. Although their indexing results were restricted to colour-texture combined methods, these methods constitute the majority of the recent methods.

Whereas some of these 1.7 millions results index supervised and semi-supervised region segmentation approaches—out of the scope of this document—, the vast amount of unsupervised methods available is still hardly manageable. We should restrict our analysis to relevant studies. However, it is hard to asses what relevant is.

Relying in our experience, in previous surveys, in the number of citations that an study has received since its publication as well as in the use of these studies for data pre-processing in image and video analysis methods, we can reduce the number of approaches. Additionally, if we restrict our analysis to recently published approaches—without ignoring classical approaches still in vogue—the range of studies is substantially narrowed.

Under the proposed organisation, we define a region segmentation approach according to its solutions to five different stages:

Pre-processing: encloses preliminary cleaning, grouping or processing of the image.

Feature extraction: defines the features used to perform the segmentation and their associated extraction process.

Local analysis: includes the comparisons performed between points in the *decision space* to derive local cues on which the segmentation relies.

Globalization: covers the extraction of holistic information in the *decision space* as well as defines the methods which operate directly on the image domain.

Regionalization: comprehends the post-processing methods applied to ensure region connectivity or contour closing as well as any other process devoted to refine the segmentation.

Not all of these stages are followed by all the existing approaches. Therefore, we propose to divide them into mandatory and optional.

In particular, all the region segmentation approaches extract a set of features to define the *decision space* Ω and perform some kind of analysis to segment the space. Optionally, some kind of pre-processing may be performed to arrange the data into a desired structure or to diminish its adherent noise. Moreover, sometimes a regionalization of the results may be required to obtain a region segmentation in agreement with the definition given in chapter 2.

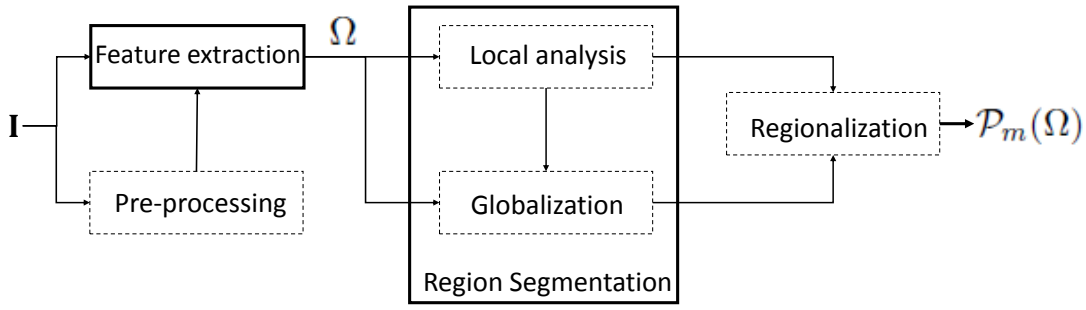


Fig. 3.2. Generic flowchart to describe region segmentation approaches. Modules in solid (dashed) lines are mandatory (optional) in a region segmentation approach.

Approaches can be also organised according to the level of processing on which the segmentation process is applied: we can distinguish: pure local—just rely on local analysis—, pure global—operate on the whole image domain—, and combined. Amongst pure local approaches we identify three principal trends: **clustering** , **region merging** and **mode-seeking**. Pure global approaches are mainly defined by an **energy minimisation** problem. Finally, in combined approaches, local information is usually represented globally through a **graph**. If this graph agglutinates inter-pixel similarities, it is the result of a **hierarchical** local analysis. On the contrary, if inter-pixel dissimilarities are instead encoded in the graph, a preliminary **contour detection** stage has been performed.

A generic flowchart to describe the processing path of region segmentation approaches is included in Figure 3.2, whereas representative approaches in each field are organised according to the defined categories in Table 3.1. The following sections are used to describe selected approaches.

3.3 (Pure) local approaches

Clustering.

In its simplest form, clustering is a spatially blind technique wherein the image data is viewed as a point cloud on the *decision space*. The core of main clustering approaches if the C-Means (or K-Means) algorithm (MacQueen et al. [1967]). The C-Means algorithm partitions a set of d-dimensional points into C clusters by minimizing an objective function. The potential of C-Means is the construction of a Voronoi tessellation of the *decision space*, i.e. a clustering of the space into sites such that the points on the boundary between two sites are equidistant from the sites representatives. C-Mean related approaches are still being used for region segmentation by incorporating spatial constraints either in the *decision space*, in the clustering process itself or by regionalization of the results. Successful schemes that rely on C-Means include Hoang et al.

	Approach	Features				Pre-processing	Local analysis		Globalization		Regionalization
		S	Y	C	T		Field	Method	Field	Method	
Local	Hoang et al. [2005]	✓		✓	✓	PCA	**Clustering**	C-Means			Cluster fusion
	Shi and Funt [2007]			✓	✓	QPCA					Refinement
	Mignotte [2008]			✓							
	Achanta et al. [2012]	✓		✓							Connectivity
	Gong et al. [2013]			✓				Fuzzy-C-Means			
	Felzenszwalb [2004]	✓	✓	✓		Smoothing	**Region-merging**	Y comparison			
	Ugarriza et al. [2009]			✓	✓			Gradient Segment.			Texture coherence
	Comaniciu and Meer [1999]	✓	✓	✓			**Mode-seeking**	MS			
	Comaniciu et al. [2001]	✓	✓	✓				Adaptive bandwidth			Mode fusion
	Comaniciu [2003]	✓	✓	✓				Bandwidth selection			
Global	Brox and Weickert [2006]	✓	✓	✓					Energy minimisation	Active contours	
	Chan et al. [2001]	✓	✓							Active contours+M-S	
	Sawatzky et al. [2013]	✓	✓							M-S	
Combined	Sharon et al. [2006]	✓	✓				**Hierarchical**	Y comp.	Graph-based	SWA	Top-down (texture)
	Alpert et al. [2012]	✓	✓		✓			Probabilistic			
	Malik et al. [2001]	✓			✓		**Contour detection**	Multi-cue comp.		Spectral cluster	Graph coarsening
	Arbelaez et al. [2011]	✓	✓	✓	✓			on circular patches			owt+ucm

Table 3.1: Proposed organisation of our selection of unsupervised region segmentation approaches. Several features can be used for the segmentation: spatial (S), Luminance (Y), Colour (C) and Texture (T). The methods used in the local analysis and in the globalization stage can be used to organise the approaches. Furthermore, by observing the empty spaces we can distinguish between pure local—the first ten approaches—pure global—the three following—and combined—the last four—. See text for details, additional examples and further acronyms definition.

[2005]; Shi and Funt [2007]; Mignotte [2008]; Achanta et al. [2012]. Let us briefly describe these studies.

In Hoang et al. [2005] , C-Means is used to illustrate the potential of a colour-texture measurement. This measurement results from the integration of the colour and texture information under a scale-space basis. The integration relies on the transformation of the energy density function implicitly captured in the image to wavelength-specific Fourier domains. Then, in these domains, colour and texture are sampled by separate Gaussian probes which are then integrated together to sample the whole transformed energy for each pixel. As multiplications with these probes are equivalent to convolutions in the spatial domain, this process can be approximated by the use of Gabor filters on the opponent colour representation of the image (Geusebroek et al. [2000]). Furthermore, logarithmic approximations are used to derive a quasi-shadow-shading invariant extraction of the Gabor filters. As the number of Gabor filters required to representatively sample the spatial, colour and texture spaces is large—60 filter responses are proposed—and as the resulting features are supposed to be highly correlated, a preliminary Principal Component Analysis (PCA) is applied on the feature space to reduce its dimensionality down to four. In order to avoid the influence of extraction noise in the PCA process, a preliminary Gaussian smoothing is applied on each response. Authors propose to apply the C-Means algorithm on the so-obtained *decision space* using a large value for C. In other words, they propose to over-segment the space in regions. Segmentation is then refined by merging

regions which centroids (in this case, the representative of the obtained sites) are similar under a Mahalanobis comparison.

The algorithm is able to moderately face texture transitions, but presents problems in the presence of uneven illumination (see results in Hoang et al. [2005] and results on the Berkeley data-set in Ilea and Whelan [2011]). In any case, the algorithm proves that the design of a *powerful decision space* may provide acceptable segmentation results even by relying in a arguably simple segmentation technique. However, the limitations of this approach are placed in the segmentation process, specially in the selection of C.

On a slightly different approach, Shi and Funt [2007] propose to use Quaternions to combine colour and texture information. The algorithm starts by extracting an orthogonal basis from the colour-textures in an image. This is achieved by sampling a set of square windows from the image and by transforming them to their Quaternion form. Then, a PCA analysis modified to operate with Quaternions (QPCA) is used to reduce the dimensionality of these samples. The authors found experimentally that reduction down to just a single dimension was enough to provide a good basis for segmentation. The C-Means algorithm is then run on the *1*-dimensional space with an empirical C=15. Several post-processing techniques are applied on the C-Means output to achieve the segmentation, including: a Gaussian spatial smoothing on the output, a region merging under a covariance-weighted measure, a spurious regions elimination and a connected component analysis.

The algorithm elegantly combines colour and texture information, but again presents an empirical selection on the number of cluster. Further experiments are required to evaluate the influence of the drastic dimensionality reduction.

Another interesting clustering method is the one described in Mignotte [2008]. In this method, the author proposes to combine C-Means results obtained when clustering an image transformed to different colour-spaces. First, the algorithm starts by generating quantised colour versions of each colour space. Then, individual C-Means processes are applied on each quantised colour space. To fuse results for these clustering processes, C-Means is applied on a feature image which comprises individual C-Means results in the shape of histograms. A version of the Bhattacharya distance is used for this C-Means. Obtained labels are then refined by merging and connected component analysis. The author admits that the algorithm is highly parameter-dependent and that a carefully parameter tuning is required to obtain good segmentations.

To end with C-Means-related approaches, we should mention Achanta et al. [2012], which aims to partition the space into superpixels. Superpixels can be understood as— usually small—regions that are somehow regular in their shape. Superpixels are intensively used in the literature due to the efficiency of their extraction processes as well as due to their conservative decisions, through which provide almost error-free over-partitions of the space.

The solution proposed in Achanta et al. [2012] has been empirically proved to outperform existing superpixel methods in nearly every respect. The algorithm is quite simple and can be understood as a set of local C-Means algorithms operating in small neighbourhoods around the representatives of the sites. These representatives are locally recomputed in each iteration. The algorithm uses an spectral-spatial combined distance for the clustering and regionalizes the converged clusters by a connected component analysis.

The algorithm just requires the setting of a single parameter: the number of desired superpixels. However, the value of this parameter severely affects the segmentation result.

Finally, let us present Gong et al. [2013] as a representative of Fuzzy-C-Means approaches. Fuzzy C-Mean has received a high amount of attention, specially from the medical image analysis research community. In its original definition, Fuzzy C-Mean is also a spatial-bind approach. Therefore, for its use in region segmentation approaches to be successful—specially in its response to noise—, spatial constraints have been included in the process. In Gong et al. [2013], the solution to overcome this problem is the introduction of a trade-off weight fuzzy factor, i.e. a weight which controls the degree of learning of each point in each iteration. The value for this weight is set through the study of the luminance distribution around a point by using a spatial kernel. The weight factor is then function of the distribution of the local information and of a local spatial constraint (hence the trade-off name) around each point in the space. Once the algorithm converges, a *defuzzification* process is applied by assigning each point to the cluster that maximizes its membership degree. The number of desired clusters is a parameter for the algorithm, and, in this case, is known a priori.

Fuzzy approaches can yield accurate segmentations if proper parameters are tuned—essentially the number of desired clusters—. However, this tuning is essentially the factor that inhibits its use for generic approaches.

Region-merging.

Region-merging approaches are initialisation dependant, which means that their final segmentation depends on how the initial seeds for merging are fixed. i.e. the first pixels on which the merging hypothesis are evaluated. Two interesting methods to overcome this problem are found in the literature: Felzenszwalb [2004] and Ugarriza et al. [2009]. Let us briefly describe them.

In Felzenszwalb [2004], authors propose to partition the space into components such that the resulting segmentation is neither too fine nor too coarse. To this aim, they first presort the pixels in a non-decreasing order according to their difference to their adjacent pixels. Then, they iteratively merge components by accounting for the internal difference and the structural shape of the regions to be merged. The algorithm is designed such that a preference for a particular region size can be included in the comparison. No especial regionalization techniques are used but a preliminary smoothing is suggested.

Under an similar scheme, in Ugarriza et al. [2009] authors present an algorithm that performs region-merging on the gradient magnitude and combine the results by texture aggregation of resulting regions. The algorithm is composed of three modules. The first module implements an edge-detection algorithm to produce an edge-map. The edge-detection is done by thresholding the gradient magnitude on the CIELab colour space. This threshold is set by means of an statistical analysis of the gradient histogram, ensuring that flat areas of the image are all included as seeds in the region-merging process. Then, the gradient threshold is progressively increased, then removing edges for each increment and producing new region-merging hypotheses. The regions are merged—simultaneously in all the space at the same time—under a colour difference premise. On a second module, authors propose to extract the entropy of the CIELab distribution around a pixel (in a 9x9 neighbourhood) to provide a compact measure of the area texture. Finally, as the colour merging process results in an over-segmentation of the image, these regions are merged by searching for texture coherence—in Mahalanobis terms—in the third module.

These approaches present interesting methods to overcome the region-merging initialisation dependency and do not require an initial estimation of the number of regions, as clustering approaches. However, they rely on some empirical similarity thresholds that also require a proper tuning.

Mode seeking.

Mode seeking is a category mainly covered by Mean-Shift (MS) approaches. MS is a non-parametric technique for data analysis. It was firstly proposed in Fukunaga and Hostetler [1975] in the scope of pattern recognition, initially oriented to the task of gradient estimation on the probability density function of the data. In Comaniciu and Meer [1999]; Christoudias et al. [2002] the technique is adapted to the task of region-segmentation. In general terms, regions are formed by grouping together pixels whose convergence points are closer to a determinate spatial and a determinate spectral quantity. These quantities, the bandwidths of the spatial and the spectral kernel, fully define the process for a given density. MS, as a generic non-parametric technique facilitates the analysis of multidimensional feature spaces with arbitrarily shaped clusters. MS has been mainly applied on luminance and colour features—albeit solutions on a texture space have been suggested (Ozden and Polat [2007])—. Usually, a regionalization technique to fuse modes is applied at the end of the process. More details can be found in chapter 4 of this document.

MS spectral bandwidth is the most relevant parameter of the algorithm. Therefore, it was natural to propose methods to set it automatically. Several efforts have been done in this subject. In Comaniciu et al. [2001] a quite complex scheme that incorporates an adaptable bandwidth term in the computation of the MS vector is proposed. However, it requires the computation of an initial bandwidth guess by a suitable plugging-rule, which, for some multi-dimensional feature

spaces is not a trivial task. In Hong et al. [2007] several schemes for improving MS operation on natural images are proposed. Among them, maybe the most relevant is the definition of a plugging-rule to estimate the bandwidth in a 3-dimensional scenario. However, the achieved background estimation is the same for all the points in the space. Therefore, whereas being the best solution in overall, this bandwidth produces an over- or under-segmentation of the image in some areas. Alternatively, in Comaniciu [2003], authors propose the computation of several MS segmentations, each one with a progressively increasing bandwidth. The optimal bandwidth for each point can be selected by searching for stability in the estimated distributions. This process requires the selection of an adequate bandwidth range, which is neither a trivial task when analysing natural images. Furthermore, the requirement of multiple segmentations severely affects the efficiency of the process.

3.4 (Pure) global approaches

In contrast to the pure-local segmentation approaches discussed up to this point, energy-based segmentation techniques aim to minimize explicit cost functions. We can classify these approaches into those that explicitly utilize edge/contour-based energy (e.g., active contours) or those that employ region-based energy to delineate different regions (e.g., Mumford-Shah formulation). These techniques usually assume that a uniform—albeit noise affected—background is present on the image and that the foreground and homogeneous objects present different characteristics to it. This configuration is uncommon in natural images. Nonetheless, these techniques are widely used in medical image applications and hence, deserve a brief mention in this chapter. We just give here a couple of intuitions about the methods; further details about energy minimisation techniques can be found in Vantaram and Saber [2012]

Active contours and Mumford-Shah functional

Active contours techniques can be divided into parametric active contours (PAC) and geometrical active contours (GAC). PAC, also named as snakes (Kass et al. [1988]), are dynamic curves that evolve based on a specific energy model until they attain a shape that best fits to an object (or to multiple objects) of interest in the scene. GAC are evolving curves which are evaluated as the level sets of a distance function in 2 dimensions. Differently from PAC, these techniques have been used to extract regions in the presence of more than two/three objects of interest (Brox and Weickert [2006]). Nonetheless, the results obtained for this method when analysing natural images are still behind those from pure-local and combined approaches.

The Mumford-Shah functional-based (M-S in Table 3.1) techniques appeared as an evolution of active contours in Chan et al. [2001], removing the dependency of the edge-based energy term of active contours. However, this technique implicitly assumes a Gaussian model version to

yield convergence. In Sawatzky et al. [2013] this assumption is eliminated by the incorporation of alternative noise models in the functional: Poison and multiplicative speckle noise.

3.5 Combined approaches/ Graph-based globalization.

We can identify two main trends in graph-based segmentation: hierarchical aggregation and contour completion.

Hierarchical aggregation.

The hierarchical region merging approach used to exemplify this category is the one proposed in Sharon et al. [2006]. It starts from pixels and extracts particular features on them. Then, these pixels are fused into bigger segments or regions according to a set of local comparisons. These regions are further fused into bigger regions until a singular region representing the whole image is achieved, then creating a hierarchical description of the image, with each level in the hierarchy representing the image under different coarseness criteria.

The merging process that leads to the hierarchical structure was firstly defined in this work and, there, was named Segmentation by Weighted Aggregation (SWA in Table 3.1).

When the whole hierarchy has been built, SWA starts from the top of the hierarchy and go down in the hierarchy by searching for texture homogeneity. Once the level at which a region fulfilling a particular homogeneity criterion is found, this region is recursively collapsed onto its *child* regions—those region on which a given region is divided in lower levels in the hierarchy—. At the end of this process, when the lowest level of the hierarchy is reached, a subset of pixels in the original image has been assigned to that region, thereby leading to the segmentation of the image.

In Alpert et al. [2012] this approach is improved by two schemes: 1) incorporating texture information in the merging process and 2) creating probability models to decide on the region merging. The so-designed algorithm substantially outperforms the results in Sharon et al. [2006].

Contour detection.

The selected contour detection approach is the one described in Arbelaez et al. [2011]. This approach relies on contour detection methods to perform the segmentation. In particular, it starts from the description of the image through several luminance, colour and texture cues extracted at different scales. Then it follows an early stage of detection, where the contour intensity of each pixel is evaluated by measuring the distribution of the extracted cues on a neighbourhood around the pixel under a set of discrete orientations (similarly to Martin et al.

[2004]). These intensities are then used to establish relations between the image pixels. These relations are encoded into an affinity matrix which describes the edge strength between any two pixels in the image—in practice, only a neighbourhood of each pixel is analysed as suggested in Malik et al. [2001]—. This matrix inherently defines and adjacency graph on which the pixels are the nodes and the edge strength between any two nodes is defined by the maximum strength in the shortest path that connects the two pixels. An spectral segmentation of the graph provides additional cues for the detection of contours. Specifically, a subset of the matrix eigenvectors is used to extract new contour cues as additional mid-level cues that integrate global information of the image. These mid-level contour cues are fused with the local cues to generate a combined strength of contour. So-obtained contours likelihoods are closed through their alignment with a Watershed transform (owt in Table 3.1) . The so-obtained boundaries are organised in a hierarchical structure under a set of strength and geometric criteria (ucm in Table 3.1).

In an earlier approach on spectral segmentation, Malik et al. [2001] authors propose to create clusters on the image graphs under the normalized-cut framework (Shi and Malik [2000]), following a two-step process. First, use a C-Means algorithm to cluster the 11 principal eigenvectors of the affinity matrix. This results in an image over-segmentation but also in a much simpler graph-representation of the image. Then, contract the simplified graph under a convergence criterion. In comparison, the approach in Arbelaez et al. [2011] yields better regions when analysing the Berkeley data-set (Martin et al. [2001]).

Finally, a more efficient alternative to the normalized-cut framework is presented in Leordeanu et al. [2012]. This alternative is named soft-segmentation. Essentially, soft-segmentation is motivated by the observation that objects in an image can be modelled by a particular—yet of indeterminate complexity—colour distribution with independence of their texture. For each image patch, this distribution is assumed to be generated by a linear combination of a finite number of colour probability distributions. Under two consecutive PCA analysis, an 8-dimensional soft-segmentation that relates every pixel in the image with every other pixel is achieved. Whereas this approximation relies on local cues but does not provide global but mid aggregation of these cues, the obtained representation is of a higher granularity of that achieved by Shi and Malik [2000], i.e. it is more influenced by noise. However, its computation is much more efficient that this one.

3.6 Evaluation of region segmentation approaches.

The exponential growth in the number of applications involving region segmentation has resulted in the creation of several data-sets for evaluation. Furthermore, among the cited papers we can recall a discussion about what is the set of evaluation statistics that better provides a quantitative measure of the segmentation quality. We will include some of this discussion here, after a

review of part of the available data-sets and—when required—of their associated evaluation methodology.

Data-sets and evaluation methodology.

Among the existing data-sets for evaluation, the Berkeley segmentation data-set Martin et al. [2001] (extended in Arbelaez et al. [2011] to conform the final BSD500 data-set) is clearly accepted as the common benchmark. The BSD500 data-set is composed of 500 natural images, with at least five user annotations each. However, the problem of the BSD500 data-set is the existence of several ground-truth annotations (one per each user) and the lack of coherence amongst the users.

Integrating this information into a rigorous evaluation framework is not an easy task. For instance, if the regions are evaluated by the overlapping of their contour with the human annotated contours, any localization error would be well tolerated. This can hardly be understood as a fair evaluation procedure as even humans localize boundaries on different pixels (see top left column in Figure 3.1).

In order to account for boundary dis-alignment, in Martin et al. [2001] authors propose to convert the detected-annotated correspondence problem into a minimum cost bipartite assignment problem, where the weight between a detected boundary pixel and an annotated boundary pixel is proportional to their relative distance in the image plane. One can then declare all boundary pixels matched beyond a given threshold to be non-hits.

To solve the multiple user problem, it has been proposed to evaluate the quality of the segmentation against each user separately. Overall results can be then obtained by averaging the statistics over the different users. Using this methodology, in order to achieve perfect recall, regions boundaries must explain all the human annotations.

Despite the *supremacy* of the BSD500 data-set we can find previous and posterior data-sets in the literature. However, we will focus in a couple of posterior data-sets. In this vein, it is worth to mention the data-set proposed in Alpert et al. [2012]. This data-set contains 200 grey-level images, half of them containing a single salient foreground object and the other half depicting 2 objects of the same type. This data-set is of main interest for the evaluation of their approach, but lacks of the variety in the BSD500 data-set and does not contain colour information. Alternatively, the Prague Texture Segmentation Data generator and Benchmark (Haindl and Mikeš [2008]) appears as a benchmark for other purposes (texture segmentation). However, as it also includes natural colour images is also a valid set for the evaluation of region segmentation approaches.

Evaluation measures.

There are two main trends to quantitatively measure the quality of region segmentation approaches: by observing the region boundaries or by considering the whole region instead.

On the one hand, the evaluation of the contour quality is usually performed under a classical *Fscore* evaluation on the ROC curve. That is if, on one side, precision (P) is measured as the ratio between the true detected boundary pixels and the total amount of boundary pixels detected, and, on the other side, recall (R) is the ratio between the true detected boundary pixels and the total amount of boundary pixels annotated, the $Fscore$ measure can be determined as:

$$Fscore_{\alpha} = \frac{PR}{\alpha R + (1-\alpha)P} \tag{3.1}$$

, where α is usually set to $\alpha = 0.5$.

On the other hand, for the evaluation of the region quality, several measures have been proposed.

The *variation of information* (VI, Meila [2005]) measures the distance between two segmentation in terms of their average conditional entropy. Being $\mathcal{P}_{n_1}$ and $\mathcal{P}_{n_2}$ two segmentations (assume that one is provided in the ground-truth) and being $H(\mathcal{P}_{m_1})$ the conditional entropy of segmentation $\mathcal{P}_{n_1}$ and $I(\mathcal{P}_{n_1}, \mathcal{P}_{n_2})$ the mutual information between the two segmentations, the VI between them is given by:

$$VI(\mathcal{P}_{n_1}, \mathcal{P}_{n_2}) = H(\mathcal{P}_{n_1}) + H(\mathcal{P}_{n_2}) - 2I(\mathcal{P}_{n_1}, \mathcal{P}_{n_2}) \tag{3.2}$$

The *rand index* (RI, Rand [1971]) operates by comparing the compatibility of assignments between pairs of elements in the segmentation. In other words, being $\mathcal{P}_{n_1}$ and $\mathcal{P}_{n_2}$ two segmentations of an image with s pixels, the RI is given by the addition between the sum of pixels with the same label in both segmentations (a) and the sum of pixels with different labels (b) , divided by the total number of pairs of pixels in the image:

$$RI(\mathcal{P}_{n_1}, \mathcal{P}_{n_2}) = \frac{a+b}{\binom{s}{2}} \tag{3.3}$$

, that is the agreement pairs divided by the total pairs.

The *segmentation covering (SC*, Everingham et al. [2008]) is usually used in pixel-classification tasks. Given two regions $\Omega^1 \in \mathcal{P}_{n_1}$ and $\Omega^2 \in \mathcal{P}_{n_2}$ their overlap is given by:

$$O(\Omega^1, \Omega^2) = \frac{|\Omega^1 \cap \Omega^2|}{|\Omega^1 \cup \Omega^2|} \tag{3.4}$$

, and the segmentation covering of segmentation $\mathcal{P}_{n_2}$ by segmentation $\mathcal{P}_{n_1}$, by :

$$SC(\mathcal{P}_{n_1} \rightarrow \mathcal{P}_{n_2}) = \frac{1}{s} \sum_{\Omega^2 \in \mathcal{P}_{n_2}} (\left|\Omega^2\right| \max_{\Omega^1 \in \mathcal{P}_{n_1}} O(\Omega^1, \Omega^2)) \tag{3.5}$$

3.7 Discussion.

Top-performing approaches are not stable along data-sets. For example, Arbelaez et al. [2011] is still the top-performing algorithm in the BSD500 data-set. Meanwhile, Alpert et al. [2012] leads the operation on the data-set they propose. Its worth to mention that Leordeanu et al. [2012], not being an explicit region segmentation method, but a contour detection method, is quite close to Arbelaez et al. [2011] performance in the task of contour detection while being much more efficient.

In this line, we have not discussed the efficiency of the methods. It is needed to say that, in general, no existing region segmentation approach is able to operate under a real-time premise. However, some approaches (Achanta et al. [2012]) are close to this requirement.

The proposed organisation covers the principal region segmentation methods currently used by researchers in several fields. From its observation emerge several open lines of research. For instance:

1. Neither clustering nor mode-seeking approaches rely on global cues.
2. Global approaches rarely operate in alternative spectral spaces rather than colour.
3. The holistic inference of combined approaches is always organised on a graph-basis and is always performed after the local-analysis, not before.

The organisation proposed in Vantaram and Saber [2012], possibly the best survey available, defines an *over-segmented* classification of existing approaches. In comparison, the organisation proposed in this chapter *under-segments* the region segmentation field, in order to provide flexibility for the classification of potential novel approaches.

3.8 Chapter conclusions.

There are multiple factors to consider when organizing region segmentation approaches and new factors and problems appear as new scenarios require the use of these approaches. The complexity of the problem has motivated excellent surveys. However, the fast evolution of the field and the huge number of applications that use region segmentation inhibits the creation of a keystone survey. In this chapter we have proposed a flexible organisation of existing approaches and exemplify this organisation by some of the most relevant approximations in each of the established categories. Let us end the chapter with a reflection: even though the proposed

organisation covers the top referenced and most-used methods in different research fields, we are sure that the proposed organisation would be, sooner rather than later, out-of-date.

Chapter 4

Mean-Shift Region segmentation based on the automatic bandwidth selection in the scale-space

This chapter proposes a region segmentation (RS) approach based on the Mean-Shift (MS) algorithm. As discussed in chapter 3, the MS bandwidth controls the aggregation criteria of MS and is the most sensible parameter in MS approaches. Therefore, an optimal selection of the bandwidth is determinant for a successful operation of MS approaches.

In this chapter, we propose to select the bandwidth by analysing the local distribution of the *decision* space. In particular, the main novelty of the proposed method is the automatic computation of a specific spectral bandwidth for each MS input sample. The so-obtained bandwidths generally increase MS convergence by adapting to the underlying data distribution as well as avoid the creation of partitions that do not have a global interpretation. The automatic bandwidth selection is achieved by a scale-space analysis of the luminance distribution of a given image.

A flowchart of the proposed method is depicted in Figure 4.1. Next sections are devoted to explain each of the algorithm's stages. Section 4.1 reviews the MS algorithm and its adaptation to RS. Section 4.2 describes the proposed method for automatic bandwidth selection in the scale-space. Section 4.3 integrates both techniques into a novel MS-RS scheme and presents a post-processing method to handle over-segmented areas. Finally, section 4.4 evaluates the proposed approach in comparison with the leading MS method in the SoA and with the most used MS approach and section 4.5 concludes the chapter.

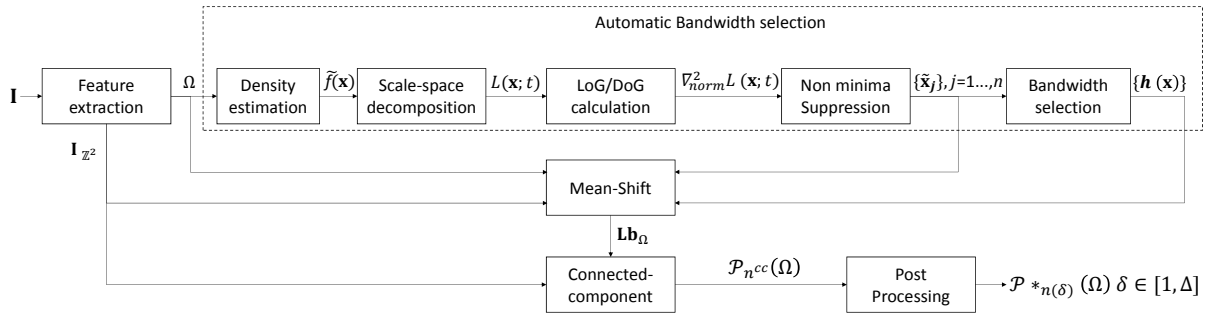


Fig. 4.1. Flowchart of the proposed MS-RS approach. For a given input image $\mathbf{I}$, the proposed method outputs a connected component RS segmentation $\mathcal{P}_{n^{cc}}(\Omega)$ and a hierarchy of post-processed segmentations $\left\{\mathcal{P} *_{n(\delta)} (\Omega)\right\}$, $\delta \in [1, \Delta]$. See text for details.

4.1 The Mean-Shift algorithm for region segmentation.

In this section, we review the Mean-Shift (MS) theory and its application for Region-Segmentation (RS). The section is organised as follows. First, an overview of MS is presented on an historical-basis, relating MS with gradient-ascend and kernel-based density estimation methods. Then, the classical MS algorithm is sketched. Next, existing solutions for the selection of MS parameters are discussed. Finally, the adaptation of the MS algorithm to RS is briefly described.

The Mean-Shift as an adaptive gradient ascend method

MS is a local non-parametric technique for data analysis. It was first introduced by Fukunaga in 1975 (Fukunaga and Hostetler [1975]) in the scope of pattern recognition as an intuitive method to estimate the gradient of the probability density function (p.d.f.) underlying a set of samples. However, it wasn't until 1995 when MS was first proposed for clustering purposes (Cheng [1995]).

In particular, in Fukunaga and Hostetler [1975] the authors select a Gaussian kernel function as a well-known differentiable kernel which satisfies a set of consistency and unbiasedness conditions for density-based estimations—to name: non-negativity, piecewise continuity and monotonicity—. Under this choice, authors observe that, given a finite set Ω of s samples in $\mathbb{R}$ and a Gaussian kernel centred at a sample $\mathbf{x}$, the mean of the displacements—or shifts, hence the name—from $\mathbf{x}$ to its neighbouring samples weighted by the kernel was proportional to the gradient of the p.d.f. evaluated at $\mathbf{x}$: $\nabla f(\mathbf{x})$.

In order to achieve a simple and manageable expression, authors opt for using a flat kernel which also fulfils the required conditions (Epanechnikov [1969]). Through such a flat kernel, $K_h(\mathbf{x})$, of bandwidth h, the expression of the mean displacement, i.e. the Mean-Shift (MS), is defined as:

$$M_{K_h}(\mathbf{x}) = \frac{1}{s} \sum_{\mathbf{x}_i \in \Omega} K_h(\mathbf{x} - \mathbf{x_i}) \tag{4.1}$$

In Cheng [1995], the author adapts this expression by introducing the concept of *sample mean* at $\mathbf{x}$—a kernel-weighted mass-centre—:

$$m_{K_h}(\mathbf{x}) = \frac{\sum\limits_{\mathbf{x}_i \in \Omega} K_h(\mathbf{x} - \mathbf{x_i})\mathbf{x}_i}{\sum\limits_{\mathbf{x}_i \in \Omega} K_h(\mathbf{x} - \mathbf{x_i})} \text{ such that }, K_h(\mathbf{x}) = \begin{cases} K(\mathbf{x}) & \|\mathbf{x}\| \leq h \\ 0 & elsewhere \end{cases} \tag{4.2}$$

The expression is valid for any kernel, $K_h(\mathbf{x})$, fulfilling the required conditions—i.e. non-flat kernels are also allowed—.

Moreover, the author proves that the MS expression can be obtained by means of the *sample mean* as:

$$M_{K_h}(\mathbf{x}) = m_{K_h}(\mathbf{x}) - \mathbf{x} \tag{4.3}$$

, which is the commonly used MS vector expression.

Furthermore, in Cheng [1995] the author explores the intuition of Fukunaga and Hostetler [1975] and establishes a condition required for the MS vector to be in the gradient direction of the underlying p.d.f. $f(\mathbf{x})$. In particular, this is fulfilled if, given two kernels $K(\mathbf{x})$ and $G(\mathbf{x})$—the former used for the MS process and the latter used on a kernel-based density estimation (KDE) of $f(\mathbf{x})$—, $G(\mathbf{x})$ is a *shadow* kernel of $K(\mathbf{x})$.

Although in Cheng [1995] a definition for the concept of *shadow* kernel is given, we prefer the one offered in Comaniciu and Meer [2002] which is focused on a particular type of kernels and would allow us to introduce their work.

Let us define $G(\mathbf{x})$ as a kernel satisfying:

$$G(\mathbf{x}) = c_g g(\|\mathbf{x}\|^2) \tag{4.4}$$

, where $g(\mathbf{x})$ is a function called the profile of $G(\mathbf{x})$ and c_g a strictly positive normalization constant which makes $G(\mathbf{x})$ integrate to one.

The kernel $G(\mathbf{x})$ is defined as a *shadow* kernel of $K(\mathbf{x})$ if their profiles $g(\mathbf{x})$ and $k(\mathbf{x})$ are related by:

$$k(\mathbf{x}) = -g'(\mathbf{x}) \tag{4.5}$$

As aforementioned, in Cheng [1995] and Comaniciu and Meer [2002] this concept of *shadow* kernel is used to prove the original intuition of Fukunaga and Hostetler [1975]. Let us reproduce here the proof to motivate the kernel selected for the proposed approach.

The KDE of $f(\mathbf{x})$ with kernel $G(\mathbf{x})$ being radially-symmetric is well-known to be expressed as:

$$\hat{f}_G(\mathbf{x}) = \frac{1}{sh} \sum_{\mathbf{x}_i \in \Omega} G\left(\frac{\mathbf{x} - \mathbf{x_i}}{h}\right) \tag{4.6}$$

, note that the bandwidth dependency is now included in the kernel argument.

Employing the profile notation, the KDE can be also expressed as:

$$\hat{f}_G(\mathbf{x}) = \frac{c_g}{sh} \sum_{\mathbf{x}_i \in \Omega} g\left(\left\|\frac{\mathbf{x} - \mathbf{x_i}}{h}\right\|^2\right) \tag{4.7}$$

, with the estimation of the gradient being computed as the gradient of the estimation as:

$$\hat{\nabla} f_G(\mathbf{x}) = \nabla \hat{f}_G(\mathbf{x}) = \frac{2c_g}{shh^2} \sum_{\mathbf{x}_i \in \Omega} (\mathbf{x} - \mathbf{x_i}) g'\left(\left\|\frac{\mathbf{x} - \mathbf{x_i}}{h}\right\|^2\right) \tag{4.8}$$

, by using equation 4.5, the expression stands for:

$$\nabla \hat{f}_G(\mathbf{x}) = \frac{2c_g}{shh^2} \sum_{\mathbf{x}_i \in \Omega} (\mathbf{x_i} - \mathbf{x}) k\left(\left\|\frac{\mathbf{x} - \mathbf{x_i}}{h}\right\|^2\right). \tag{4.9}$$

Operating to adapt the equation to the MS vector expression:

$$\nabla \hat{f}_G(\mathbf{x}) = \frac{2c_g}{shh^2} \left[\sum_{\mathbf{x}_i \in \Omega} k\left(\left\|\frac{\mathbf{x} - \mathbf{x_i}}{h}\right\|^2\right)\right] \left[\frac{\sum\limits_{\mathbf{x}_i \in \Omega} k\left(\left\|\frac{\mathbf{x} - \mathbf{x_i}}{h}\right\|^2\right) \mathbf{x}_i}{\sum\limits_{\mathbf{x}_i \in \Omega} k\left(\left\|\frac{\mathbf{x} - \mathbf{x_i}}{h}\right\|^2\right)} - \mathbf{x}\right] \tag{4.10}$$

, where we have reached the desired expression.

The left part of equation 4.10 is proportional to a KDE with kernel $K(\mathbf{x})$ whereas the right part is the MS vector (see equations 4.2 and 4.3). The relation can be easily observed if we replace in equation 4.10 the equivalent of equation 4.7 for $K(\mathbf{x})$ and use equation 4.3 to isolate the MS term:

$$\nabla \hat{f}_G(\mathbf{x}) = \frac{2c_g}{c_k h^2} \hat{f}_K(\mathbf{x}) M_K(\mathbf{x}) \tag{4.11}$$

and,

$$M_K(\mathbf{x}) = \frac{1}{2} h^2 c \frac{\nabla \hat{f}_G(\mathbf{x})}{\hat{f}_K(\mathbf{x})} \tag{4.12}$$

, where we have combined both normalisations constants into c.

In equation 4.12, the MS vector with kernel $K(\mathbf{x})$—here represented through its profile

Algorithm 4.1 The mean-shift algorithm

For a finite set of samples Ω, and given a sample $\mathbf{x}$, a kernel $K(\mathbf{x})$ and a bandwidth h

1. Compute the sample mean $m_K(\mathbf{x})$ by equation 4.2
2. Compute the mean-shift vector $M_K(\mathbf{x})$ by equation 4.3
3. Translate the centre of kernel $K(\mathbf{x})$ to $m_K(\mathbf{x})$.
4. Repeat 1,2,3 until $M_K(\mathbf{x}) \cong 0$,

$k(\mathbf{x})$—is proportional to the gradient of the estimated function with kernel $G(\mathbf{x})$—hence it is *aligned* with it—. Therefore, as intuited by Fukunaga and Hostetler [1975], the MS vector points towards the maximum variation of the p.d.f around $\mathbf{x}$.

There is another interesting implication of equation 4.12. The MS vector, $M_K(\mathbf{x})$, is related with the gradient of the KDE obtained by using the kernel $G(\mathbf{x})$, $\nabla \hat{f}_G(\mathbf{x})$, by a *normalisation* function $\hat{f}_K(\mathbf{x})$. This function is the KDE of $f(\mathbf{x})$ obtained with the MS kernel: $\hat{f}_K(\mathbf{x})$. This implies that the MS vector becomes small in high populated areas—near the local maxima of the distribution, where $\hat{f}_K(\mathbf{x})$ is high—. On the contrary the MS vector gets larger in low populated areas—associated with small values of $\hat{f}_K(\mathbf{x})$—. For this reason, MS has been claimed to be and adaptive gradient ascend method.

The Mean-Shift algorithm

The MS algorithm, defined as in Algorithm 4.1, is guaranteed to converge (Comaniciu and Meer [2002]) to a point where the estimated function $\hat{f}_G(\mathbf{x})$ has zero gradient—a point named as a stationary point—. In other words, the algorithm converges to a local maximum in the surroundings of $\mathbf{x}$. The algorithm can also converge to a local minimum—which is also a stationary point—, albeit, due to its gradient-ascend nature, this is only possible if $\mathbf{x}$ is initially placed on it—.

Additionally, as the convergence of the MS algorithm is guaranteed, it automatically defines a clustering procedure. If the MS process is performed on every sample in $\Omega = \{\mathbf{x}_1, ..., \mathbf{x}_s\}$ several samples will converge to the same local maximum, a mode $\tilde{\mathbf{x}_j}$ in the distribution. The MS trajectories followed by these samples will visit common locations—common values for $m_K(\mathbf{x})$—in some iteration. These locations drive the MS process to converge to $\tilde{\mathbf{x}_j}$. These locations are known as the *basin of attraction* of the mode $\tilde{\mathbf{x}_j}$.

All the samples that converge to $\tilde{\mathbf{x}_j}$ can be well represented by $\tilde{\mathbf{x}_j}$. In other words, these samples can be grouped into the same cluster Ω_j, with $\tilde{\mathbf{x}_j}$ being the mode of such—arbitrary-shaped—cluster. This is due to the fact that, as explained in chapter 2, the mode for a region is unique, and the *basin of attraction* of $\tilde{\mathbf{x}_j}$ establish that $\tilde{\mathbf{x}_j}$ is the only stationary point within

some open sphere on the data, with the kernel $K(\mathbf{x})$, and the bandwidth, h, defining such sphere.

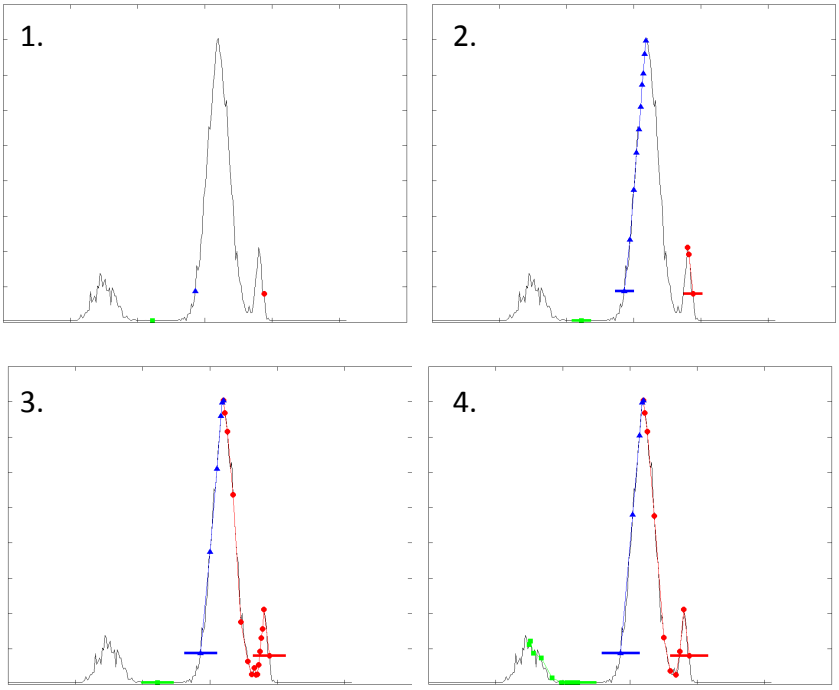


Fig. 4.2. Bandwidth effect in Mean-Shift. From left to right and up to down. (1.) Given an unknown p.d.f. in $\mathbb{R}$—here depicted in black—with three clear modes, the MS algorithm (Algorithm 4.1) is applied on three samples—here represented by three different coloured shapes—. Using these samples for the initialisation of three MS processes, each one of these process is expected to converge to a different mode. (2.) For a given h the MS process applied on the red and the blue samples converges to the nearby modes. However, the green sample remains in a plateau area. (3.) On a given increment of h—the value of h is here represented by the length of a solid line on the input samples—the MS for the red sample converges to a different mode, hence the other mode is now unreachable by the process. Nevertheless, the green sample remains in the flat area of the distribution. (4.) Finally, if h is further increased, the MS for the green sample ends to converge to its closest mode on the distribution. In conclusion, there is not a global value for h that associates each sample with its *expected* mode. Additionally, note how the number of steps in the MS trajectories (represented by the number of coloured points over the p.d.f.) is reduced as h is increased, indicating that the convergence of the algorithm is faster.

Kernel selection

Regarding to the kernel selection, if the aim is to increase the quality of the KDE process in equations 4.6 and 4.7, the Epanechnikov kernel (Epanechnikov [1969]) has been proven to minimize the asymptotic approximation of the mean square error (AMISE) between the density and its estimate (Scott [2015]). The Epanechnikov kernel is expressed in $\mathbb{R}$ as:

$$G(\mathbf{x}) = \begin{cases} \frac{3}{4}(1 - \|\mathbf{x}\|^2) & , \|\mathbf{x}\| \leq 1 \\ 0 & , \|\mathbf{x}\| > 1 \end{cases} \tag{4.13}$$

, whereas the Epanechnikov is a *shadow* of the flat kernel—aside for normalisation—:

$$K(\mathbf{x}) = \begin{cases} 1 & , \|\mathbf{x}\| \leq 1 \\ 0 & , \|\mathbf{x}\| > 1 \end{cases} \tag{4.14}$$

Therefore, the flat kernel is an optimal one to use in the MS process under the given AMISE minimization aim, and under the constraint imposed by Cheng [1995] and Comaniciu and Meer [2002] that have been explained earlier in this section.

Bandwidth selection

The bandwidth selection is more complex. Let us first discuss the problematic involved in the selection of the bandwidth and then review existing approximations to set its value.

From the kernel and the MS vector definitions it is clear that the bandwidth h drives the process. If a large value for h is selected, the algorithm is prone to converge fast as more samples in the distribution are used to compute the MS vector. However, if a narrow mode is close to another mode—closer than h for the flat kernel—the narrow mode would be occluded by this other mode and none sample will converge to it. Instead, if a small value for h is selected, the algorithm will converge slower, but narrow modes are prone to be detected. However, with a small h, MS can get stuck in plateau areas which are also zero gradient, then creating non-local-maxima modes.

A graphical example of these situations is depicted in Figure 4.2. In the example, it is shown how increasing the value of h is required for some samples to allow the MS process to *escape* from plateau areas of the underlying distribution. However, increasing the value of h also implies that narrow modes are progressively occluded by closer wider modes, hence inhibiting MS to converge to the narrow modes. The example also illustrates how even by selecting a *good* overall estimate for h—we will explain what *good* is in this context later on, at this point let us define it as what can be *expected* from the shape of the distribution—it may not be *good* for all the samples at the same time.

Whereas in many cases the bandwidth h is selected manually according to the data, several strategies have been proposed to set a *good* value for h automatically.

Statistically-motivated methods search for a bandwidth value that returns the best bias-variance trade-off of the estimator, i.e. minimises the AMISE. There are several heuristic strategies to obtain this value in the unidimensional case, which are known as *plug-in-rules* (Comaniciu and Meer [1999, 2002]). However, the dependency of these rules with the curvature

of the p.d.f hinders its application in higher dimensional spaces (Hong et al. [2007]). Furthermore, we have observed in Figure 4.2 that a solely bandwidth may not be enough to provide a *desired* behaviour of MS on all the samples.

Stability-searching methods repeat MS with several h hypotheses and select the optimal value as the mid bandwidth of the largest bandwidth range over which MS produces the same number of clusters. This scheme—which best example is Comaniciu [2003]—provides a generic method for the selection of h which is able to be applied on practically any scenario if the initial range of hypotheses is adequately selected. However, it implies successive computations of the MS process and requires the bandwidth range to be highly sampled, i.e. consecutive tested bandwidths should be similar as, otherwise, the repeatability of the number of clusters is not ensured. Additionally, this scheme allows to set different values of h for different samples in the distribution; at the expense of substantially increasing the computational cost of the algorithm.

Quality-guided methods rely on an objective function that determines the *goodness* of the resulting clusters. These functions are typically related with the balance between inter- and intra-cluster variance (Kaufman and Rousseeuw [2009]). Whereas these techniques achieve compact and distinctive clusters, they also require the swapping of several bandwidth hypotheses as well as the definition of a suitable objective function that may change for each faced task.

The higher-level-dependent methods rely on the use of stages of analysis placed on a higher hierarchy in the semantic pyramid (see chapter 2) to solve a problem which is task-dependent most of the times. Whereas this is a natural way to perform—as MS is usually applied as an early stage of processing—feed-back techniques should be designed *ad hoc* for each task.

A local-adaptive-bandwidth method was proposed in Comaniciu et al. [2001]. In particular, the authors propose an elegant solution to adapt the bandwidth locally, introducing the adaptive scheme in the core of the MS process. The process relies on the adaptive gradient idea that MS provides naturally and force the local bandwidth to maximise the value of the MS vector on each sample. Let us illustrate this by an intuition: in plateau areas of the distribution, the MS vector is high as the density is low—see equation 4.12—. If h is increased until a maximum of $M_K(\mathbf{x})$ is reached, such h value provides that all the samples in the plateau area are *contained* in the kernel, hence, the immediately next h will include a sample out of the flat area, allowing MS to *escape* from it. However, the process is complex to reproduce—which partially explains its lack of use in posterior applications—and requires the prior estimation of a pilot density under an initial bandwidth hypothesis.

Finally, more heuristic rules—for instance, making h proportional to the average distance of each sample to its k^{th} nearest neighbour in Ω—have been also proposed. These rules also allow the definition of a different bandwidth for each sample. However, they are highly dependent of the faced task and also require the choice of the number of considered neighbours.

In this chapter we propose a simple scheme to estimate the bandwidth *a priori* for every

sample. It relies on a prior globalization of the problem and also constrains the number of reachable samples that the MS process can achieve. However, before describing the process in detail let us first explain how the MS process is applied for RS.

Region-segmentation via Mean-Shift

MS can be applied directly for RS if no spatial-constraints are required. Let Ω be the set of 1-dimensional descriptions of the image $\mathbf{I}$, e.g. the luminance values of each pixel; hence s—the number of samples—is the number of pixels in the image. Let $\mathbf{x}$ be one of the samples in Ω and let $K(\mathbf{x})$ and h being properly selected, the Algorithm 4.1 can be straightly used.

For each $\mathbf{x}$, MS converges to a mode $\tilde{\mathbf{x}_j}$, $j \in [1, n]$, where n is the number of modes. By assigning to each sample $\mathbf{x}$ the identifier of the mode—i.e. j— a RS of $\mathbf{I}$ into n regions according to Ω is obtained:

$$\mathcal{P}_n(\Omega) = \left\{ (\Omega_1, ..., \Omega_n) \ : \ \Omega = \bigcup_{j=1}^{n} \Omega_j \text{ and } \Omega_j \cap \Omega_k = \emptyset \text{ for all } j \neq k \right\} \tag{4.15}$$

Note that this is the same expression presented in equation 2.1. Furthermore, a region representative—the mode $\tilde{\mathbf{x}_j}$ (see chapter 2)—is automatically obtained for each region as the point of convergence of the MS algorithm—in fact we have been using the mode terminology up to this point, as it is the common one used by MS methods—.

As explained in chapter 2, a proper RS method is required to group connected samples. This is not ensured in this case as several unconnected points in the image lattice might converge to the same mode.

As generally known, luminance values in $\mathbf{I}$ are organised in a 2-dimensional lattice $\mathbf{I}_{\mathbb{Z}^2}$ of d-dimensional vectors—$d = 1$ for a luminance-driven analysis, which is the one described in this chapter—. The spatial coordinates in the lattice of the luminance sample $\mathbf{x}$ are here represented by the 2-dimensional vector $\mathbf{p} = (u, v)$.

In order to incorporate this spatial information into the MS process, the MS kernel is usually defined as the product of two radially symmetric kernels (Comaniciu and Meer [2002]). One of the kernels is devoted to account for spatial information $K_{\mathbf{p}}(\mathbf{p})$ and the other one deals with the luminance values, $K_\Omega(\mathbf{x})$. Representing the kernels by their profiles $k_{\mathbf{p}}$ and k_Ω, the sample mean for sample $\mathbf{x}$ is obtained as:

$$m_{h_{\mathbf{p}},h_\Omega}(\mathbf{x}, \mathbf{p}) = \frac{\sum\limits_{\mathbf{x}_i \in \Omega} k_{\mathbf{p}} \left(\left\| \frac{\mathbf{p}-\mathbf{p_i}}{h_{\mathbf{p}}} \right\|^2 \right) k_\Omega \left(\left\| \frac{\mathbf{x}-\mathbf{x_i}}{h_\Omega} \right\|^2 \right) \mathbf{x}_i}{\sum\limits_{\mathbf{x}_i \in \Omega} k_{\mathbf{p}} \left(\left\| \frac{\mathbf{p}-\mathbf{p_i}}{h_{\mathbf{p}}} \right\|^2 \right) k_\Omega \left(\left\| \frac{\mathbf{x}-\mathbf{x_i}}{h_\Omega} \right\|^2 \right)} \tag{4.16}$$

, where $h_{\mathbf{p}}$ is the spatial bandwidth and h_Ω the luminance bandwidth, and the MS vector

stands as:

$$M_{h_{\mathbf{p}},h_\Omega}(\mathbf{x},\mathbf{p}) = m_{h_{\mathbf{p}},h_\Omega}(\mathbf{x},\mathbf{p}) - \mathbf{x} \tag{4.17}$$

Note that if $K_{\mathbf{p}}(\mathbf{p})$ and $K_\Omega(\mathbf{x})$ are chosen to be flat kernels—as *shadows* of Epanechnikov kernels, for the reasons previously described—the spatial kernel acts as a sample selector on which to apply the luminance kernel.

In essence, the spatial bandwidth $h_{\mathbf{p}}$ indicates how big is the spatial neighbourhood on which to apply MS for a given $\mathbf{x}$ placed at spatial position $\mathbf{p}$. Although this does not explicitly ensure connectivity, it naturally restricts the analysis to a subspace of the luminance distribution, hence including local constraints in the MS process and reinforcing spatial continuity of the clusters.

Let us briefly discuss the advantages and disadvantages of this spatially-constrained MS respect to the unconstrained scheme.

On one hand, in an hypothetical spatially unconstrained MS case, the grouping criteria—defined by the luminance kernel $K_\Omega(\mathbf{x})$ and the luminance bandwidth h_Ω—may lead to consider evidences from unconnected samples with a similar luminance value—i.e. close samples in the luminance distribution but not necessarily close in the image lattice—. This will bias the MS process towards global cues of the distribution with independence of the local luminance distribution.

On the other hand, in the spatial constrained MS, global modes on the distribution may not be included in the local neighbourhood defined by $h_{\mathbf{p}}$; hence, the MS process might be prone to converge to non-global modes. This has an over-partitioning effect in the RS, as more modes than those *existing* in the global distribution might be obtained. This is usually solved by a later stage of analysis, through a mode fusion scheme. Spatial adjacent clusters which modes are closer than h_Ω are fused into a single cluster. However, this scheme just solves the problem partially, as clusters unconnected with global-mode-representative clusters will be still associated to non-global modes.

4.2 Scale-space for MS bandwidth selection.

In this section, we propose a simple scheme which, taking advantage of the special characteristics of digital images, is able to automatically select a bandwidth for each sample $\mathbf{x}$. Whereas MS and scale-space are closely related techniques which have been combined before (Nocedal and Wright [2006]). This is, to our knowledge, the first proposal to select the MS bandwidth via the scale-space theory. The scheme is based on a prior detection of the global modes in the scale-space. Next subsections are organised as follows. First the scale-space theory is reviewed. Then, its application for mode detection in the luminance *decision* space is summarised. Next, a novel scheme for Non minima Suppression is presented and finally, the bandwidth selection scheme is described.

Scale-space decomposition and associated derivatives

Given a discrete signal $f(\mathbf{x}) : \mathbb{Z} \to \mathbb{R}$, its scale-space decomposition (Lindeberg [1993]) is the family of functions $L(\mathbf{x}; t) : \mathbb{Z} \times \mathbb{R}_+ \to \mathbb{R}$ such that $L(\mathbf{x}; 0) = f(\mathbf{x}), \mathbf{x} \in \mathbb{Z}$ and, for $t > 0$, $t \in \mathbb{R}_+$:

$$L(\mathbf{x}; t) = \sum_{\mathbf{y}=-\infty}^{\infty} g(\mathbf{y}; t) f(\mathbf{x} - \mathbf{y}) \tag{4.18}$$

, or:

$$L(\mathbf{x}; t) = g(\mathbf{x}; t) * f(\mathbf{x}) \tag{4.19}$$

, where t is the scale parameter and,

$$g(\mathbf{x}; t) = \frac{1}{\sqrt{2\pi t}} e^{-\mathbf{x}^2/2t} \tag{4.20}$$

, is. a Gaussian kernel profile.

The selection of the Gaussian kernel is motivated by several premises—see Lindeberg [1993] for details—. For instance, due to its semi-group or cascade-application property:

$$g(\mathbf{y}; t_1) * g(\mathbf{y}; t_2) = g(\mathbf{y}; t_2 + t_1) \tag{4.21}$$

, it is verified that:

$$L(\mathbf{x}; t_2) = g(\mathbf{y}; t_2 - t_1) * L(\mathbf{x}; t_1),\ t_2 > t_1 \tag{4.22}$$

In practice, this entails that if $t_1 = \sigma$, $t_2 = 2\sigma$ and $t_\iota = \iota\sigma, \forall \iota > 0$, the same Gaussian kernel profile $g(\mathbf{x}; \sigma)$ can be used along the process:

$$L(\mathbf{x}; t_\iota) = g(\mathbf{x}; \sigma) * L(\mathbf{x}; t_{\iota-1}) \tag{4.23}$$

Additionally—also due to the election of the Gaussian kernel—it is shown that the Laplacian of Gaussian (LoG), ∇^2, of each $L(\mathbf{x}; t_\iota)$ is proportional to the Difference of Gaussian (DoG) between consecutive scales:

$$\nabla^2_{norm} L(\mathbf{x}; t_\iota) = \frac{t_\iota}{(t_\iota - t_{\iota-1})} \left[L(\mathbf{x}; t_\iota) - L(\mathbf{x}; t_{\iota-1}) \right] \tag{4.24}$$

, where *norm* stands for scale-normalised.

The LoG—and hence the DoG—returns local minima (maxima) in the scale-space surface $\mathbb{Z} \times \mathbb{R}_+$ for local maxima (minima) of the function $f(\mathbf{x})$. Therefore, focusing only on the local maxima—aka the modes—of $f(\mathbf{x})$, these can be obtained, together with the scale at which they are produced, by:

$$(\tilde{\mathbf{x}},\tilde{t}) = \underset{(\mathbf{x},t)}{argminlocal}(\nabla^2_{norm} L(\mathbf{x};t)) \tag{4.25}$$

Luminance mode detection in the scale-space

Let $\hat{f}(\mathbf{x})$ be a pilot distribution—i.e., an estimate of the real $f(\mathbf{x})$—of the *decision* space Ω. The estimation of $\hat{f}(\mathbf{x})$ usually requires the selection of a kernel and a bandwidth; then, the quality of the estimation—its *divergence* respect to the real estimation—may be affected by these selections.

The set Ω of luminance data of a digital image $\mathbf{I}$ can, differently, be well represented in a constrained discrete space $\Omega = \{\mathbf{x}\} \subset \mathbb{Z}$,$\mathbf{x} \in [0, sup(\Omega)]$, where sup($\Omega$) is the supreme of the set Ω and, usually, sup(Ω) = 255 —or 100 depending on the decision space—.

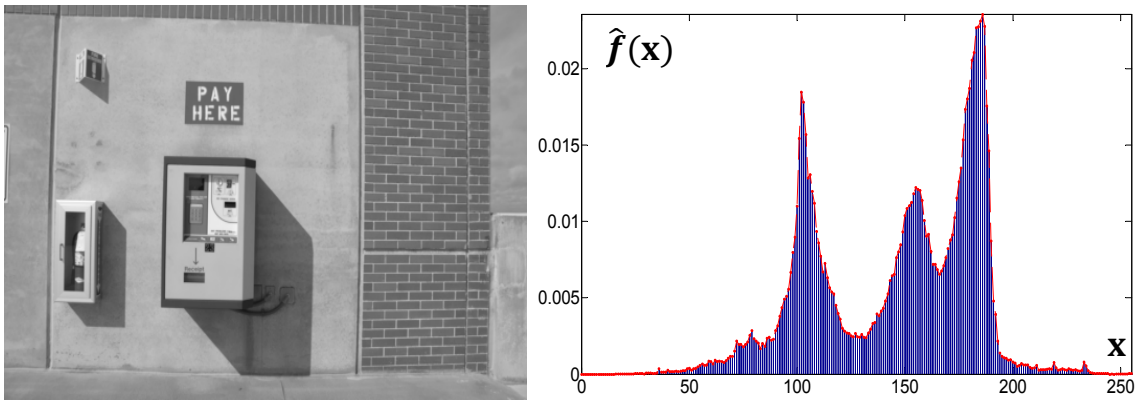


Fig. 4.3. A Luminance image (left) and its probability mass function (right). The histogram will be substituted by the envelope (red plot) in next figures for visualisation purposes.

In this case, $f(\mathbf{x})$ is a probability mass function (p.m.f.)—not a p.d.f.—and its estimation $\hat{f}(\mathbf{x})$ can be done without losing any information nor introducing any bias by: (1) computing an histogram of bin size 1 on the data in Ω on the whole range $[0, \sup(\Omega)]$ and (2) dividing the amount of samples falling in each bin by the total number of pixels in the image. This is, in general, not possible for real-valued—continuous—signals. However, the process is also applicable to other discrete *decision* spaces, e.g. the RGB-colour space. An example of a p.m.f. is included in Figure 4.3.

Once $\hat{f}(\mathbf{x})$ has been estimated, we can obtain its scale-space decomposition $L(\mathbf{x};t)$ for a particular range of scales by applying equation 4.23. Then, we can approximate the LoG of the decomposition at each scale $L(\mathbf{x};t_\iota)$ by equation 4.24 and, finally we can localise the local maxima by means of equation 4.25.

Figure 4.4 includes an example of the scale-space decomposition $L(\mathbf{x};t)$ of the p.m.f. $\hat{f}(\mathbf{x})$ represented in Figure 4.3 with $\sigma = 1$ and 100 scales. In order to provide the reader with an overall intuition of what the scale-space decomposition is, we have included three different representations: a 3-dimensional plot of the discrete samples—where the envelope of $\hat{f}(\mathbf{x})$ is also depicted in red to ease visualisation—, a 2-dimensional plot of selected scales and a matrix-like representation of the scale-space. This matrix-like representation is built by the stacking of the scale-space functions $L(\mathbf{x};t)$, i.e. the ι^{th} row of the matrix corresponds to $L(\mathbf{x};t_\iota)$.

Similarly, Figure 4.5 includes the same three visualisation for the DoG $\nabla^2_{norm}L(\mathbf{x};t)$. In the matrix-like representation—obtained under the same stacking process on the $\nabla^2_{norm}L(\mathbf{x};t)$ functions—it is clear that maxima of $L(\mathbf{x};t)$—and hence of $\hat{f}(\mathbf{x})$—are related with minima of $\nabla^2_{norm}L(\mathbf{x};t)$.

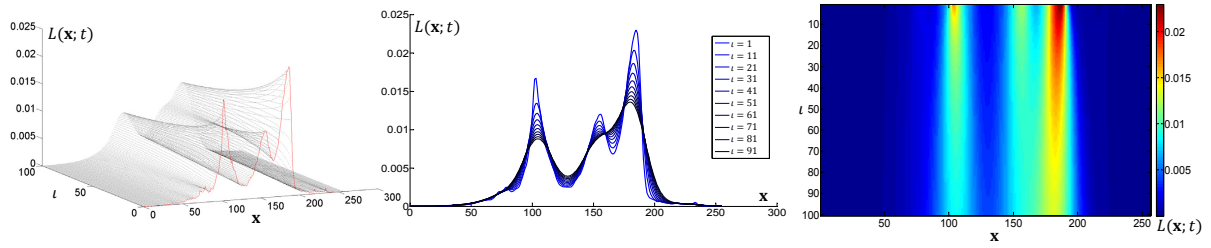


Fig. 4.4. Three representations of the scale-space decomposition of the p.m.f $\hat{f}(\mathbf{x})$ depicted in Figure 4.3. Left column: 3-dimensional plot of $L(\mathbf{x};t)$ samples; observe how the p.m.f. is progressively smoothed. Middle column: 2-dimensional plot of selected scales, see also how the Gaussian effect progressively merges local maxima. Right column: matrix representation of $L(\mathbf{x};t)$, can be seen as a zenithal view of the 3-dimensional plot. Note how the intensity of the maxima decreases with the scale, due to the smoothing effect.

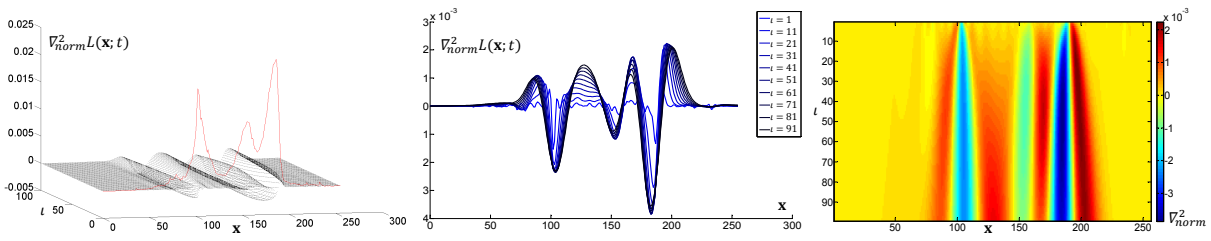


Fig. 4.5. Three representations of the DoG $\nabla^2_{norm}L(\mathbf{x};t)$ on the scale-space decomposition $L(\mathbf{x};t)$ of the p.m.f. $\hat{f}(\mathbf{x})$ depicted in Figure 4.3. Left column: 3-dimensional plot of $\nabla^2_{norm}L(\mathbf{x};t)$ samples; observe how the DoG remains almost flat for non local extrema of the p.m.f. Middle column: 2-dimensional plot of selected scales, see also how the Gaussian effect progressively merge local minima of the DoG. Right column: matrix representation of $\nabla^2_{norm}L(\mathbf{x};t)$, can be understood as a zenithal view of the 3-dimensional plot. Note how the minimum intensity of the local minima is reached at intermediate scales—not at the last scale—.

Whereas the scale-space decomposition and the DoG calculation are well-founded processes that have been recursively explored in the literature, the last step, the localisation of local minima described by equation 4.25, is the least explored and the most problematic stage. Next section is devoted to describe our solution to this problem.

Non-minimum suppression

The Gaussian kernel has the property of non-enhancing of local extrema, i.e. local maxima can not get higher after convolution and local minima can not get lower. This can be observed in the middle diagram of the scale-space representation in Figure 4.4. There, the lighter blue curves are always higher—or at much equal—than darker blue curves in the surroundings of the local maxima of $\hat{f}(\mathbf{x})$. The curves configuration is exactly the other way around in the surroundings of local minima.

Due to the scheme used to approximate the LoG—see equation 4.24—these configurations imply that, in the DoG, the local maxima of $\hat{f}(\mathbf{x})$ are always related with negative values of $\nabla^2_{norm}L(\mathbf{x};t)$. Similarly, the local minima of $\hat{f}(\mathbf{x})$ are indicated by positive values of $\nabla^2_{norm}L(\mathbf{x};t)$. For instance, this effect is evident in the middle diagram of the DoG representation in Figure 4.5.

Our solution for Non minimum Suppression (NmS) builds on this fact. Let us define the *minima blobs* as areas in the scale-space surface which extent is defined by the negative values of the DoG. In order to identify them, we can use the matrix-like representation of the DoG and isolate the *minima blobs* by removing the positive samples in the matrix. The effect of this process can be observed by comparing b. and c. in Figure 4.6.

Let us focus in the *minima blobs* areas rather than in the DoG values these contain. A connected-component analysis (see chapter 2) of the so-built matrix can be used to identify the *blobs* (see d. in Figure 4.6). However, as *minima blobs* are extracted on the last scale, information of *blobs* shaped at earlier scales may be lost. The principal cause of this problem is that low modes which are close to higher modes may be occluded by the progressive Gaussian smoothing.

We propose to handle this problem by tracking *minima blobs* in the scale-space surface. For this purpose, we will use the matrix-like representation of the DoG. The tracking procedure starts on the first row—first scale—of the matrix representation of the DoG and labels *blobs* progressively. If, at a given scale, a *blob* is fused with another—i.e. the number of *blob*s decreases respect to that of the previous scale—, a new label is set and the prior result is conserved. The process continues until all the active areas of the matrix have been inspected. The effect of the *blob* tracking procedure is the relabelling of the *minima blobs.* In Figure 4.6, the relabelling process starts from d. and obtains e.

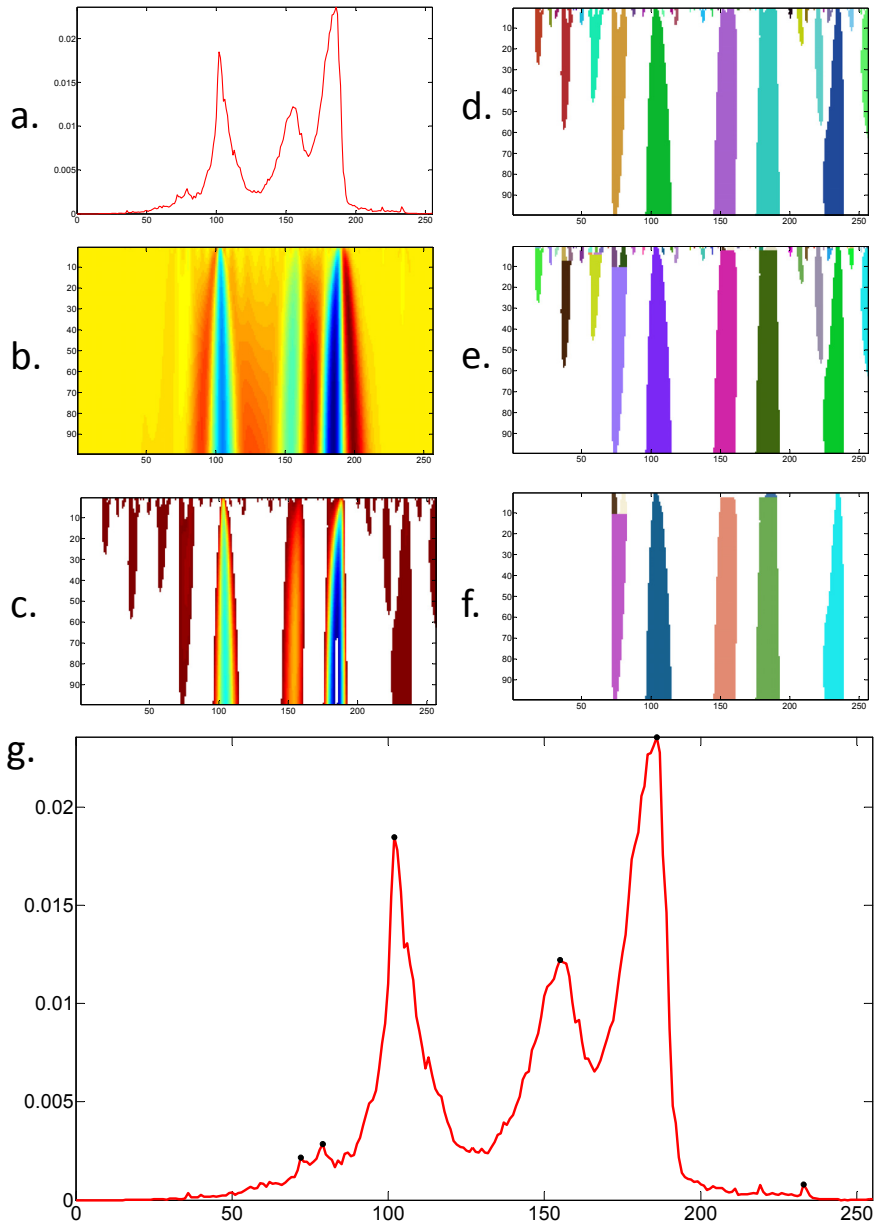


Fig. 4.6. (a.) p.m.f envelope estimation $\hat{f}(\mathbf{x})$ of luminance image in Figure 4.3. (b.) The DoG in its matrix-like representation. (c.) The negative part of the DoG: *minima blobs* are automatically detected. (d.) A connected component analysis labels each *minima blob*—each *blob* is here identified by a random colour—. However, *blob* merging in the scale-space may imply mode losing. (e.) Relabelled *minima blobs*: blobs are tracked along the scale to identify *blob* merging situations—which occur at low scales in this example—see Figure 4.7 for additional examples. (f.) *minima blobs* associated with small absolute values of DoG—in this case, of absolute value lower than $th_{\nabla^2} = 10^{-5}$—have been discarded to illustrate the threshold effect. (g.) Each remaining *minima blob* defines a mode in $\hat{f}(\mathbf{x})$, here represented by black dots on the p.m.f. envelope. Note that the blob tracking strategy prevents the merging of the two left modes.

Each *minima blob* encompasses a local area of the DoG surface. The extraction of the local minimum on each of these areas results in the identification of a mode $\tilde{\mathbf{x}}_j$ in $\hat{f}(\mathbf{x})$ and of the scale t_ι at which it is most significant. In the example these are the Cartesian coordinates of the local minima in the scale-space surface.

Inaccuracies of the scale-space estimation and image noise may create spurious modes. In order to avoid the detection of these modes, we decided to select only those *minima blobs* that have an associated minimum DoG absolute value lower than a given threshold th_{∇^2}. This threshold has a relevant effect in the segmentation results. We evaluate the sensitivity of the process to this threshold in section 4.3 .

The result of the NmS process for the p.m.f of Figure 4.3 is included in part (g.) of Figure 4.5. Additional examples are depicted in Figure 4.7. In the light of these figures, we can stand that, in overall, the method is able to yield accurate and generalist relevant-mode detection without (miss)detecting a significant number of non-relevant modes.

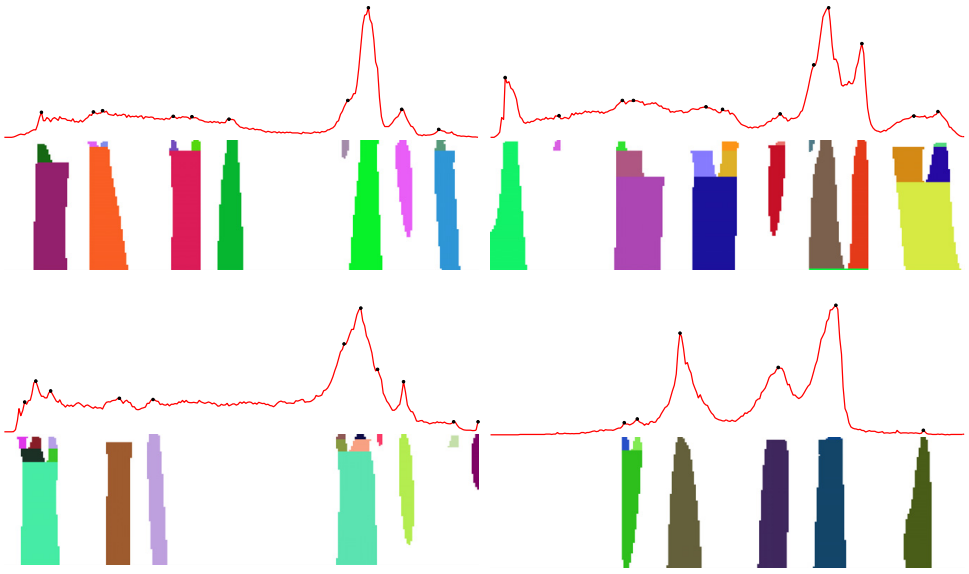


Fig. 4.7. Additional examples of Non minimum Suppression in the scale-space. Each sub-figure includes the envelope estimation $\hat{f}(\mathbf{x})$ with modes found by the proposed solution (top) and the relabelled *minima blobs* used to generate those modes (bottom)—each *blob* is here identified by a random colour—. *Minima blobs* associated with small absolute values of DoG—in this case, of absolute value lower than $th_{\nabla^2} = 10^{-5}$—have been discarded and removed from the graphic to illustrate the threshold effect.

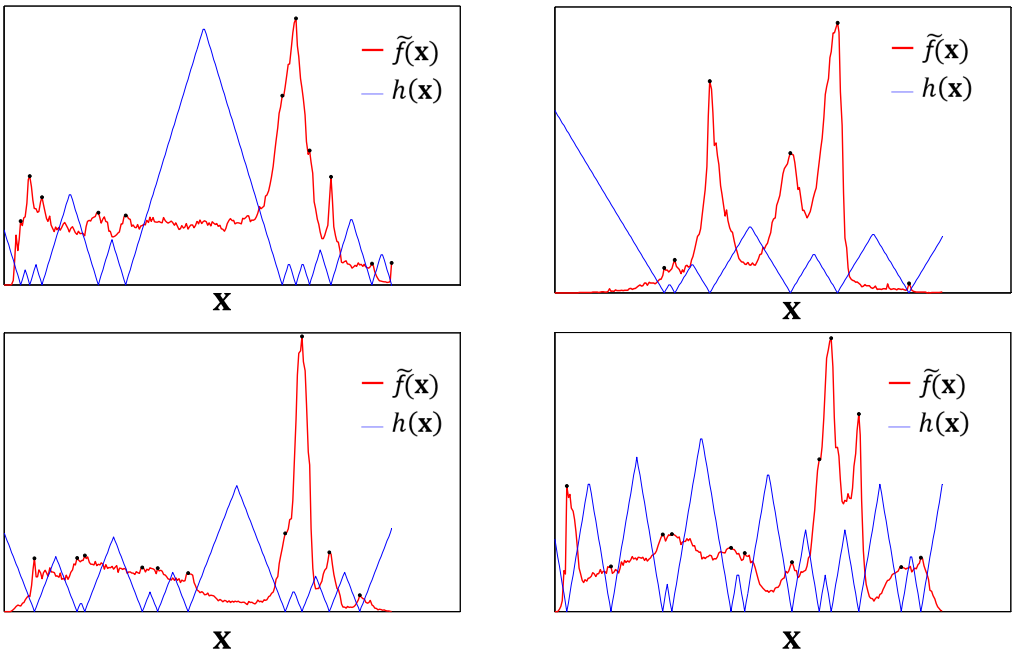


Fig. 4.8. Examples of the bandwidth selection process. In each sub-figure the red plot represents the p.m.f. envelope, detected local maxima are indicated by black dots and bandwidth is represented here as a continuous blue signal to ease figure visualization. Bandwidth value has been normalized on the p.m.f. range to allow their joint representation. Note how in plateau areas the bandwidth reaches their maxima values.

MS bandwidth selection

We aim to assign to every sample $\mathbf{x}$ a proper bandwidth $h_\Omega(\mathbf{x})$ that generally increases MS convergence as well as impedes MS blockage in plateau areas of the p.m.f. To this aim, we make use of the mode detection in the scale-space.

From the scale-space theory (Lindeberg [1993]) it is known that the scale at which each mode is best detected—local minima of the LoG/DoG—is related with the *width* of the detected modes. In particular, the response of the LoG reaches a minimum at scale $\tilde{t}$ for modes covering a range $\{\mathbf{y}\}_j$ such that $, \sup(\{\mathbf{y}\}_j) - \inf(\{\mathbf{y}\}_j) \cong 2\tilde{t}$, where $\inf(\{\mathbf{y}\}_j)$ is the infimum of the set.

Therefore, the mode detection implicitly defines a suitable MS bandwidth—twice the standard deviation of the Gaussian at the scale of detection—for all the samples in the set $\{\mathbf{y}\}_j$ of each mode $\tilde{\mathbf{x}_j}$. However, samples not assigned to any mode range won't be covered by this scheme. To face this problem, we propose a general solution to cover all the samples.

The proposed mode detection scheme samples the p.m.f.. Aside from mode detection, the mode location scheme implicitly defines plateau areas—those between two modes—on the p.m.f. With this in mind, and in order to generally reduce the number of convergence steps of MS and to avoid the stagnation of the MS process in plateau areas, we define the bandwidth $h_\Omega(\mathbf{x})$ for

each sample $\mathbf{x}$ as:

$$h_{\Omega}(\mathbf{x}) = \min_{j} \|\mathbf{x} - \tilde{\mathbf{x}_j}\| \tag{4.26}$$

, i.e. $h_{\Omega}(\mathbf{x})$ is the minimum value that will allow a global MS process to consider for every $\mathbf{x}$ at least a p.m.f. mode in the calculation of the MS vector at the first iteration. The effect of this bandwidth selection process can be inspected in Figure 4.8.

4.3 Proposed luminance-based region-segmentation approach.

This section builds on the two previous sections and presents the proposed MS-based RS (MS-RS) approach. First, the proposed solution is motivated via the discussion on two common MS problems. Next, the proposed algorithm is presented. Finally, its design limitations are discussed, leading to a post-processing solution to overcome them—or at least to minimise their impact on the final RS—.

Motivation of the proposed solution

After mode detection, a naive scheme to bypass the MS operation would be to assign each sample to the cluster represented by its closer mode in the *decision* space. However, this would be a purely global approach that would: (i) ignore the real distribution between the sample and its closer modes, and (ii) ignore the local p.m.f. distribution around each sample. Let us call these problems *miss-location* and *localisation*. A graphical sketch of these problems is included in Figure 4.9. We will use this Figure as a guide to describe these problems, disregard this scheme and motivate the one proposed.

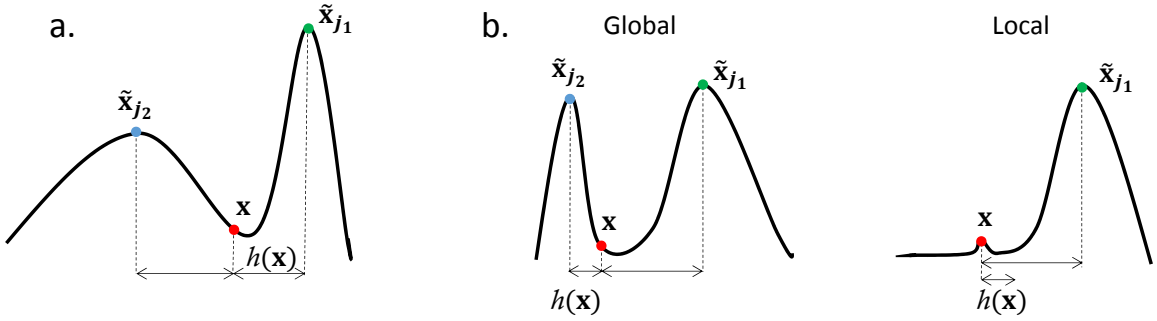


Fig. 4.9. Bandwidth selection problems. a. *miss-location*, b. *localisation*. See text for details.

miss-location (Figure 4.9 a.). A sample $\mathbf{x}$ is placed inside the range of a mode $\tilde{\mathbf{x}}_{j_2}$. However, it is closer in the *decision* space to a mode $\tilde{\mathbf{x}}_{j_1}$. The naive global clustering on the scale-space would assign $\mathbf{x}$ to the cluster represented by $\tilde{\mathbf{x}}_{j_1}$. A MS process on mode $\mathbf{x}$, even with bandwidth

$h(\mathbf{x})$ might have a *chance* to assign $\mathbf{x}$ to its proper mode $\tilde{\mathbf{x}}_{j_2}$. However this assignment would depend on the local distribution, which leads us to the next problem.

localisation (Figure 4.9 b.). A sample $\mathbf{x}$ is placed inside the range of a mode $\tilde{\mathbf{x}}_{j_2}$ (left part of b.). However, in its spatial neighbourhood—defined by $h_{\mathbf{p}}$, see equation 4.16—most of the samples shaping this mode could be out of the spatial range; samples from a nearby mode $\tilde{\mathbf{x}}_{j_1}$ are, instead, included (right part of b.). The designed bandwidth mechanism would avoid a MS process on $\mathbf{x}$ converge to $\tilde{\mathbf{x}}_{j_1}$. Instead, it would converge to a new shaped local mode around $\mathbf{x}$. If this problem is repeated for several samples, the MS process would result in the creation of several local modes that do not correspond to *existing* modes in the p.m.f.

On one hand, in order to partially handle *miss-location* problems we propose to apply the MS algorithm locally on each sample under the bandwidth selected in the scale-space. On the other hand, to correct *localisation* problems, we force the MS process to converge to one of the modes detected in the scale-space, so, at the end of the RS no new modes are created. Note that this last scheme also helps to diminish the dependency of the RS to the spatial bandwidth parameter. If the spatial kernel is wide enough to incorporate at least some evidences from nearby modes, the proposed RS method converges to these modes by the *forcing* scheme. In our experiments we observe that a spatial bandwidth of $h_{\mathbf{p}} = 10$ resulted in wide-enough kernels for the majority of the analysed images.

Proposed MS-RS algorithm.

A flowchart of the MS-RS algorithm is included in Figure 4.1 whereas the algorithm is sketched in Algorithm 4.2. Most of the algorithm stages have been already described. Nevertheless, we summarise the whole algorithm here for completeness.

Given an image $\mathbf{I}$, in an early stage of the algorithm the luminance, Ω, and the spatial information, $\mathbf{I}_{\mathbb{Z}^2}$, are extracted. The luminance information can be represented by a set of samples Ω, with each luminance sample $\mathbf{x}_i$ having a spatial coordinate associated $\mathbf{p}_i$ in $\mathbb{Z}^2$. The p.m.f. of the luminance data, $\hat{f}(\mathbf{x})$, is obtained by first constructing an histogram of the luminance data with bins of length 1 in the range $[0, 255]$ and then dividing the number of samples falling in each bin by the total number of pixels in $\mathbf{I}$.

An scale-space decomposition of the so-obtained $\hat{f}(\mathbf{x})$ is carried out with 100 scales and a Gaussian kernel of $\sigma = 1$. In general, a lower number of scales would provide good results in the analysis of natural images. However, we established 100 as a conservative solution to be able to detect extraordinary wide modes. The variance of the kernel has been selected equal to the bin size, so that every plausible mode can be observed—in discrete spaces, minimum width is 1—.

The LoG is estimated on the scale-space decomposition by means of the subtraction of consecutive scale-space functions (DoG). *Minima blobs* are shaped by accounting only for negative values of the DoG. Through the proposed NmS scheme a set of modes $\{\tilde{\mathbf{x}_j}\}, j = 1, ..., n$ is ob-

Algorithm 4.2 Proposed Mean-Shift RS algorithm

The MS-RS algorithm.
Input: a finite set of discrete-valued samples Ω extracted from an image $\mathbf{I}$.
Output: a RS $\mathcal{P}_{n^{cc}}(\Omega)$ of $\mathbf{I}$ in terms of Ω.

1. Estimate the p.m.f. of the samples in Ω.
2. Extract the modes of the p.m.f. through scale-space analysis (equations 4.23, 4.24 and 4.25) and the proposed NmS technique.
3. Select the optimal bandwidth $h(\mathbf{x})$ for each sample $\mathbf{x}$ by equation 4.26.
4. Being $K_p(\mathbf{p})$ and $K_\Omega(\mathbf{p})$ flat kernels in the shape of equation 4.14 and $h_\mathbf{p}$ a desired spatial bandwidth, do:

for each pair $\mathbf{x}$, $\mathbf{p}$:

i. Compute the sample mean $m_{h_\mathbf{p},h_\Omega}(\mathbf{x})$ by equation 4.16

ii. Compute the mean-shift vector $M_{h_\mathbf{p},h_\Omega}(\mathbf{x})$ by equation 4.17

iii. Translate the centre of kernel $K_\Omega(\mathbf{p})$ to $m_{h_\mathbf{p},h_\Omega}(\mathbf{x})$.

iv. Repeat i., ii., iii. until $M_{h_\mathbf{p},h_\Omega}(\mathbf{x}) \cong 0$.

v. Force $m_{h_\mathbf{p},h_\Omega}(\mathbf{x}) \leftarrow \tilde{\mathbf{x}}j$, $\hat{j} = \underset{j}{argmin}(\left\| m_{h_\mathbf{p},h_\Omega}(\mathbf{x}) - \tilde{\mathbf{x}}_j \right\|)$.

vi. Assign $\mathbf{Lb_\Omega}(\mathbf{p}) \leftarrow \hat{j}$

end for.
5. Obtain $\mathbf{Lb}_{\mathbf{\Omega},\mathbb{Z}^\mathbf{2}}$ and, hence, the RS: $\mathcal{P}_{n^{cc}}(\Omega)$, by a connected-component analysis of $\mathbf{Lb_\Omega}$.

tained. The bandwidth for each sample is set as its distance in Ω to its nearest mode (through equation 4.26).

A spatial constrained MS process guided by a spatial and a spectral flat kernel is then performed on each sample through equations 4.16 and 4.17. The MS process is forced to end in a mode detected in the scale-space. Therefore, the number of regions obtained by this process equals the number of modes detected in the scale-space. However, these regions may not represent connected components in the image plane and hence, a connected component analysis is applied on the regions (see chapter 2).

Sensitivity analysis

In order to study the sensitivity of the approach to the spectral threshold parameter th_{∇^2} we propose to evaluate the proposed method on the training images in the Berkeley segmentation data-set (Martin et al. [2001]; Arbelaez et al. [2011]). This set is composed of a total of 200 natural images with a minimum of five human annotations of region boundaries per image.

We follow the matching strategy proposed in Martin et al. [2001]; Arbelaez et al. [2011]

to evaluate the goodness of the method for the task of boundary detection (see chapter 3). However, we noticed that an evaluation just driven by global statistics—i.e. considering only the overall detection statistics on the complete training data-set—may not be representative of the real performance of a given method. In particular, so-extracted statistical figures may be strongly biased by results obtained for images on which a high number of boundaries have been annotated, whereas the influence in the statistics of the results obtained for images with a low number of boundaries may be hindered. Figure 4.10 illustrates this problem. The global statistics—which values would be discussed later on—suggest that the best performance (in F-Score terms) is achieved close to the last threshold analysed (close to the end of each graph). Studying the operation limits of the method, i.e. the statistical area on which results vary when considering per-image results, we observe that Precision figures are low for all the images (observe the best Precision curve) operating with such threshold values. In contrast, and in the light of the Figure 4.10, operating with intermediate values of th_{∇^2} seems to provide better statistical figures for at least one image in the data-set.

In order to measure the average performance of the method but considering the operation on each image; we propose to compute local statistics per analysed image and then average these statistics. Through this scheme, the influence of the number of annotated boundaries in the overall evaluation may be decreased. Figure 4.11 depicts so-extracted statistics. Observe the strong differences in the end of this graphic when compared with Figure 4.10. This behaviour suggests that there are a very small number of images with a high number of contours annotated, which is the case. Furthermore, note that the expected better operation for intermediate values of th_{∇^2} is now clear from the plot.

In order to increase the number of images on which the methods performs at its best achievable operation, we opt for this last scheme and select the optimal threshold on the peak of the averaged F-Score: $th^*_{\nabla^2} = 1.4388 \cdot 10^{-5}$ (indicated by a black line on the plot). The rest of the experiments in this chapter are all extracted by using this value ($th^*_{\nabla^2}$) as the threshold for the mode detection in the DoG of the luminance image scale-space decomposition and, as aforementioned, a spatial bandwidth $h_{\mathbf{p}} = 10$.

Regarding the statistics values themselves, the designed method is able to detect the majority of the annotated boundaries (averaged recall rates over 80% for intermediate values of th_{∇^2}) but operating under very low averaged precision rates (under 40% with independence of the value of th_{∇^2}). This is a clear indicator of over-segmentation. The image is divided into regions such that a minority of their boundaries are well-aligned with the annotated boundaries in the image; however, the majority of the boundaries have not been considered relevant for any of the users responsible of the annotations. In our opinion, the main cause for the over-segmentation is the inability of the proposed scheme to properly handle textured areas. In these areas, the pixel luminance varies strongly between adjacent pixels. The designed scheme

is highly sensible to luminance transitions. In textured areas these are present almost between every pair of adjacent pixels. Consequently, in these areas, almost every pixel is assigned to a different region. Next section is devoted to describe a colour-based post-processing mechanism to reduce—to some degree—the influence of textured areas in the designed method by incorporating colour information. Chapter 5 enhances this solution by also incorporating texture descriptions to the process.

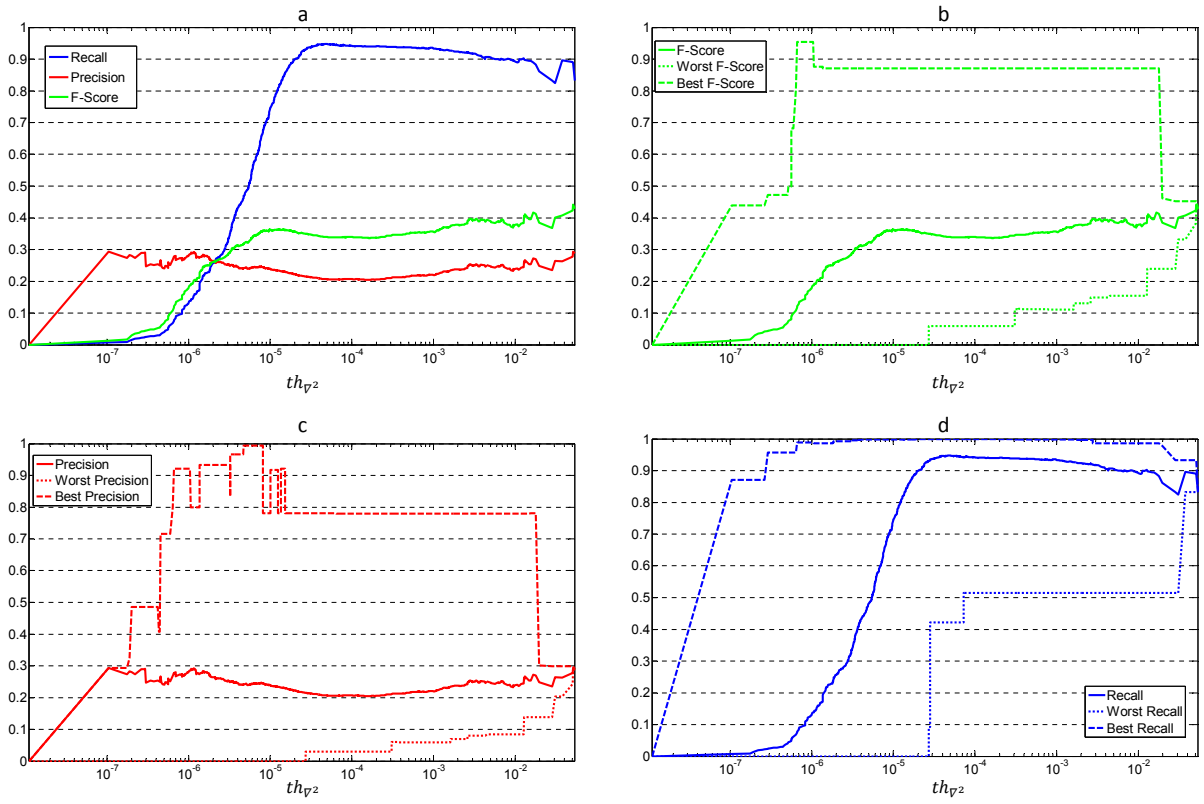


Fig. 4.10. Sensitivity analysis on the training set of the Berkeley Dataset (overall statistics on the set). All the graphics have been plotted on a logarithmic scale on the values of th_{∇^2} to ease visualisation. (a) Recall, Precision and F-Score curves for different values of th_{∇^2}. (b) Global F-score and F-score operation area (between best and worst F-score per training image). (c) Global Precision and Precision operation area (between best and worst Precision per training image). (d) Global Recall and Recall operation area (between best and worst Recall per training image).

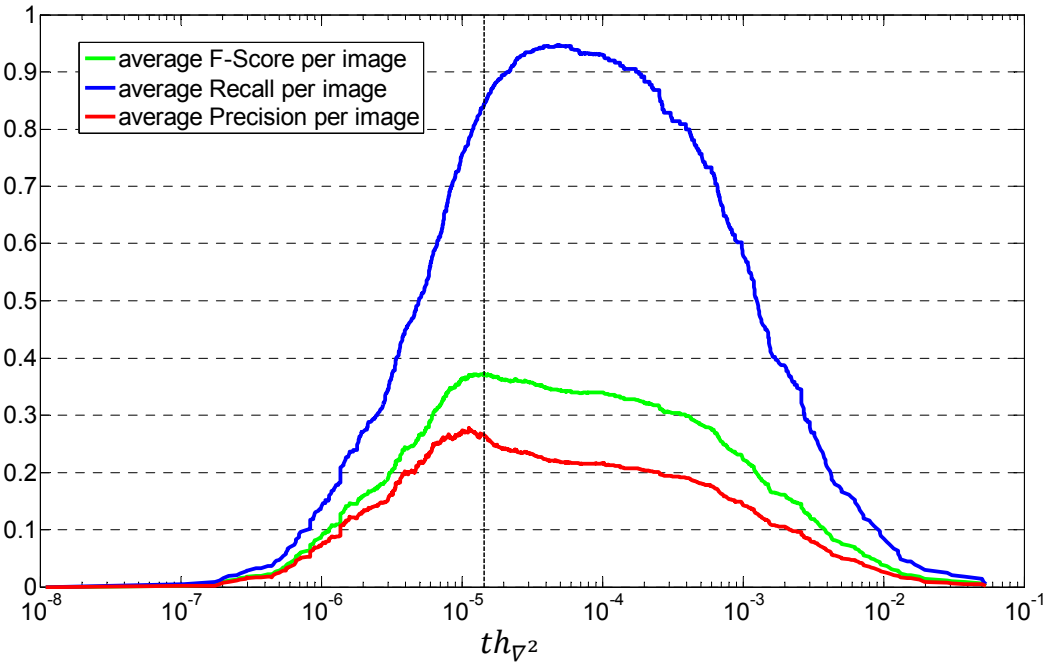


Fig. 4.11. Sensitivity analysis on the training set of the Berkeley Dataset (average statistics per image). Average Recall, Precision and F-Score curves per image for different values of th_{∇^2}. The graphic has been plotted on a logarithmic scale on the values of th_{∇^2} to ease visualisation.

Post-processing.

Inspired by the hierarchical methodology of successful state-of-the-art approaches (Sharon et al. [2006]; Alpert et al. [2012]; Arbelaez et al. [2011]) we propose to refine the luminance-based results of the proposed MS-RS approach by a hierarchical region-merging scheme. The scheme is sketched in Algorithm 4.3. The proposed merging strategy relies on three basics:

1. **Merging on the region adjacency graph.** The use of a region adjacency graph on which the candidates pairs for merging are defined.

2. **Building the merging hierarchy.** The creation of a hierarchy of merging hypotheses on which colour differences are quantised in levels.

3. **A sampling strategy to threshold distances for several images.** The unification of the distances obtained for the RS of different images, such that a generic merging hierarchy to cover all the images can be defined.

Next paragraphs describe each of these basics in detail.

Algorithm 4.3 Proposed hierarchical region-merging post-processing algorithm based on CIE-Lab colour informations.

Start from the result of the MS-RS: $\mathcal{P}_{n^{cc}}(\Omega)$; a RS composed of n^{cc} connected-component regions.

1. Build a region adjacency graph RAG on which the nodes are the regions and the edges connect each region Ω_j with all of its adjacent regions $\{\Omega_k\}$ in the image lattice.
2. Weight each edge in the graph (Ω_j, Ω_k) by a local-variability based distance between the two involved regions: $d(\Omega_j, \Omega_k)$.
3. Quantise the statistical distribution of the set of distances in the RAG, $\mathbf{d} = \{d\}_{RAG}$ by studying its $\delta\%$ percentiles.
4. Start from $\delta = 1$
   (a) while $\delta \leq \delta_T$ do
   (b) merge $(\Omega_k \leftarrow \Omega_j)$ the pair of regions which associated distance $d(\Omega_j, \Omega_k)$ is the highest amongst those lower than $\mathbf{d}_{\left(\frac{\delta \cdot P \cdot |\mathbf{d}|}{100}\right)}$.
   (c) update the distances associated to Ω_j and the edges in the RAG associated to Ω_k.
   (d) repeat (b) and (c) until no new pairs can be merged.
   (e) output $\mathcal{P} *_{n(\delta)} (\Omega)$, an $\delta-$merged version of $\mathcal{P}_{n^{cc}}(\Omega)$
   (f) $\delta \leftarrow \delta + 1$ and back to (a).

Merging on the region adjacency graph

We propose to first segment the luminance image through the proposed MS-RS scheme configured with the parameters derived from the analyses in the previous sections. Then, we construct the Region Adjacency Graph (RAG) by using region as nodes and by connecting in the graph the regions that share a common boundary. Each connection in the RAG represents a candidate pair of regions for merging.

We propose to quantify the likelihood of each pair to be merged by the distance between the representatives of these regions. In particular, we opt for the computation of the CIEDE00 (d_{00}) distance between the CIE-Lab mode vectors extracted as the median (individually computed per colour channel) of each of the involved regions. This metric is preferred to the $l^2 - norm$ due to its superior behaviour in measuring changes for small colour differences (Habekost [2013]).

A pair of regions $\{\Omega_j, \Omega_k\}$ connected in the RAG are fused if:

$$d(\Omega_j, \Omega_k) = d_{00}(\mathbf{Lab}(\Omega_j), \mathbf{Lab}(\Omega_k)) \leq th_{E_{00}} \tag{4.27}$$

, where $\mathbf{Lab}(\Omega_j)$ is the CIE-Lab colour vector representative of region Ω_j and $th_{d_{00}}$ is a threshold on the colour difference which value is studied later in this section.

As explained in chapter 3, given a particular value for $th_{d_{00}}$ the region merging process may present a couple of inconsistencies that should be clarified. Let us illustrate these via an example (sketched in Figure 4.12).

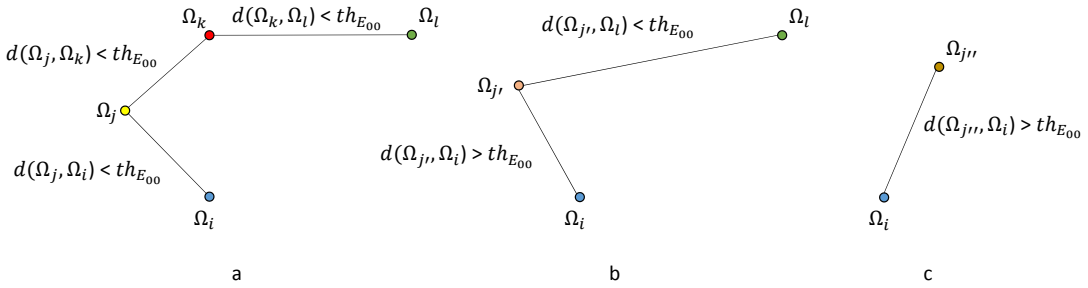


Fig. 4.12. Region merging problematic. We start with three plausible region merging situations (a). A given region-merging order (in this case merging first Ω_j and Ω_k) may change the merging scenario (b). For a given $th_{E_{00}}$, the merging process continues until all the connected regions fulfil the condition imposed by $th_{E_{00}}$ (c).

Let region Ω_j be connected with two regions Ω_k and Ω_i. Let be $d(\Omega_j, \Omega_k)$ and $d(\Omega_j, \Omega_i)$ both lower than $th_{d_{00}}$. If the region Ω_j is merged with Ω_k , these fusion creates a new region $\Omega_{j'}$, which new CIE-Lab colour vector: $\mathbf{Lab}(\Omega_{j'})$, results in a comparison $d(\Omega_{j'}, \Omega_i)$ greater than $th_{d_{00}}$. Moreover, the new region $\Omega_{j'}$ will be connected to all the regions connected to Ω_j and to all the regions connected to Ω_k. The distance between $\mathbf{Lab}(\Omega_{j'})$ and $\mathbf{Lab}(\Omega_l)$ also fulfils

the threshold condition; hence, this pair should also be merged, producing a new region $\Omega_{j''}$. Therefore:

i. The region merging order might determine the final result.

ii. The region merging process cannot be defined simply between every pair of adjacent regions but considering the whole *flow* on the graph (as illustrated by the max-flow min-cut theorem in Elias et al. [1956]; Boykov [2001]).

To face these problems, we opt for a simple scheme. Given a threshold $th_{d_{00}}$ and a RAG, we propose to merge first the regions which colour vectors convey the highest distance which is lower than $th_{d_{00}}$ (facing i.) and iterate under this basis until all the candidate pairs fulfilling the threshold have been merged (facing ii.).

Building the merging hierarchy.

The use of different values for $th_{d_{00}}$ would result in different merging results. In fact, as $th_{d_{00}}$ is increased, the more regions would be merged and the more simpler (composed of a lower number of regions) the segmentation would be. In our hierarchical scheme, we define increasing threshold values $th_{d_{00}}(\delta)$. Each of this threshold values define a level in a hierarchy of region partitions $\left\{\mathcal{P} *_{n(\delta)} (\Omega)\right\}$, $\delta \in [1, \delta_T] \subset \mathbb{Z}$, with δ_T being the total number of thresholds explored such that: $th_{d_{00}}(1) < th_{d_{00}}(2) < ... < th_{d00}(\delta_T - 1) < th_{d_{00}}(\delta_T)$.

A sampling strategy to threshold distances for several images

Due to the different spectral properties of natural images, region connections in different images are prone to convey different colour distances. The colour distances for each image will vary on a different distance range and will be differently distributed than the colour distances in another image. This turns the selection of a generic set of threshold values to cover all the image scenarios a highly complex—if not infeasible—task.

Distance normalisation—i.e. dividing distance values by the maximum distance in the RAG—will place the distances extracted on several images on a common distance range (as it is done in Arbelaez et al. [2011]). However, normalisation will not solve the different-distribution problem, which still precludes the establishment of a common sampling strategy on the threshold values. We propose to rely on the use of percentiles to define a sampling scheme for $th_{d_{00}}$. To this aim, being $\mathbf{d} = \{d\}_{RAG}$ the ascending-ordered set of colour distances in the RAG of each analysed image, we set:

$$th_{d_{00}}(\delta) = \mathbf{d}_{\left(\frac{\delta \cdot P \cdot |\mathbf{d}|}{100}\right)} \tag{4.28}$$

method	Boundary							Region					
	Recall		Precision		F-Score		Area	SC (GT)		RI		VI	
	ODS	OID	ODS	OID	ODS	OID	under Curve	ODS	OID	ODS	OID	ODS	OID
$\mathcal{P}_{n^{cc}}(\Omega)$ $(th^*_{\nabla^2})$	0.94	0.94	0.21	0.21	0.34	0.34	-	0.16	0.16	0.73	0.73	9.03	9.03
$\mathcal{P}^*_{n(\delta)}(\Omega)$ $(\delta \in [1,100])$	0.66	0.74	0.41	0.45	0.51	0.56	0.41	0.44	0.56	0.76	0.80	2.46	2.11
Arbelaez et al. [2011]	-		-		0.73	0.76	0.73	0.59	0.65	0.81	0.85	1.65	1.47

Table 4.1: Quantitative results for the test images of the BSD500 data-set. ODS: Optimal operation point for the whole test set of images. OID: Optimal operation point computed individually for each image. The proposed approaches present worse statistics that Arbelaez et al. [2011] in every of the quantitative statistic evaluated (see section 3.6 in chapter 3 and Arbelaez et al. [2011] for details). The problem is the low precision of the proposed approach—even after region-merging—which is an indicator of over-segmentation.

, were $\delta \cdot P$ % indicates a particular percentile, P controls the sampling, $|\mathbf{d}|$ is the cardinality of the set $\mathbf{d}$, and $\mathbf{d}_{\left(\frac{\delta \cdot P \cdot |\mathbf{d}|}{100}\right)}$ is the $\left(\frac{\delta \cdot P \cdot |\mathbf{d}|}{100}\right) -$ ordered statistic of the distances in $\mathbf{d}$.

In our experiments, we set $P = 1$ and hence, set $\delta_T = 100$. For instance, for $\delta = 50$, $th_{d_{00}}(50) = \mathbf{d}_{\left(\frac{50|\mathbf{d}|}{100}\right)}$. In words, the threshold is equal to the distance which is higher than the 50% of the distances in $\mathbf{d}$, i.e. it is equal to the median of the set of distances.

By this scheme we define the thresholds as function of the image colour transitions. Hence, with independence of the image, this process quantise the colour distances in a fixed number of levels.

4.4 Experimental results

Experiment description

We quantitatively evaluate the proposed approach on the test images in the Berkeley Dataset (BSD500, Martin et al. [2001]; Arbelaez et al. [2011]). Results are included and compared with the leading algorithm in the state-of-the-art in Table 4.1. The proposed MS method is evaluated under the optimal configuration according to the sensitivity analysis ($th^*_{\nabla^2}$). The post-processing scheme is evaluated for all the $\delta \in [1, 100]$ hypotheses. Statistics are returned both for the problem of boundary location and region construction (see details in Arbelaez et al. [2011]).

Images in the data-set present highly textured areas which neither the proposed MS-RS nor the proposed merging scheme are able to cope with. Nonetheless, and in order to provide a qualitative evaluation of the proposed system operating on a scenario best suited for its characteristics—e.g. a less textured one—, we present qualitative results on the data-set proposed in Guo et al. [2013].

Qualitative results of the proposed MS-RS algorithm before $[\mathcal{P}_{n^{cc}}(\Omega)]$ and after post-processing refinement are included in the fifth and sixth rows of Figures 4.13 , 4.14, 4.15, 4.16 and 4.17. In these Figures we qualitatively compare our algorithm with the EDISON system [1]. The EDISON system operates on colour images and combines MS (Comaniciu and Meer [2002]) and edge detection (Meer and Georgescu [2001]) to perform synergistic image segmentation (Christoudias et al. [2002]). To perform a comparison on even grounds, we have set equal spatial bandwidth for both systems $h_{\mathbf{p}} = 10$ and run the EDISON system with three different spectral bandwidths $h_{\Omega} = \{5, 10, 20\}$.

Discussion

In the light of the results, we think that the proposed algorithm is able to obtain as accurate representations of the scene as the EDISON system operating only on the luminance channel. Furthermore, obtained regions present boundaries which are tighter to the scene contours—without relying in a contour detection mechanism—. The algorithm is able to respect the fine details of the images as well as to handle moderate noise in large homogeneous areas. Moreover, as desired, there is no need to set the bandwidth parameter.

In essence, we have replaced this parameter by a threshold on the DoG: th_{∇^2}. The dependence of the system to this parameter is relatively high (see Figures 4.10 and 4.11). Whether this threshold is easier to set that the MS spectral bandwidth is leaved to the reader's opinion. In our opinion, this parameter controls the number of final luminance modes. This affects partially to non-mode samples that are placed in plateau areas flanked by these modes, but does not affect the rest of the samples.

Regarding the post-processing mechanism, whereas it is effective for refining problematic areas on some images, it is unable to handle strongly textured regions, where final RS keeps oversegmented (see 4.17 and Table 4.1). In general, this approach respects MS-RS segmentations in non-texture areas and respects the fine details of the image if these are continuous in luminance; observe this situation in Figure 4.18. On the contrary, the fine detail is lost sometimes if the luminance regions are not distinctive enough for their surrounding regions (represented by strong colour differences)—see failure cases of the post-processing method in Figure 4.19—.

The main problem of the proposed approach is its high sensitivity to luminance transitions. Note that, whereas the bandwidth selection scheme severely reduces the number of modes on which MS can converge, the construction of connected-component regions substantially increases the final number of regions; hence, producing over-segmentation. These regions are highly different to their neighbouring regions in textured areas. The proposed post-processing scheme does not property handle this problem as it is unable to discriminate between problematic and not-problematic regions. This is clear by observing Table 4.1. The region-merging post-

[1] http://coewww.rutgers.edu/riul/research/code/EDISON/

processing improves the Precision (merge problematic regions) of the MS-RS a 95%. However, this is achieved at the expense of reducing MS-RS Recall (eliminate correct contours) a 27%. In overall (according the F-Score value), the region-merging process improves the MS-RS a 50%, i.e. from 0.34 to 0.51, still far from the best existing operation (0.73 by Arbelaez et al. [2011]). Region statistics are also improved in similar terms, but the overall operation is still well below the leading approach in the field.

4.5 Chapter conclusions.

In this chapter we have proposed a novel method to automatically select the spectral bandwidth for each input sample in a MS scheme by an a priori analysis in the scale-space decomposition of the input data. We have first reviewed the MS and scale-space theories to later present their combination: MS-RS. We have then qualitatively evaluated MS-RS and discovered that it was unable to handle textured areas. A region-merging post-processing method to face these problematic areas was presented. In overall, the method is able to outperform or at least equal the operation of the most popular MS method in the literature on not-highly textured images. However, the proposed solution for handling textured areas is still unable to face all the problems in these areas; hence, the achieved RS is still too over-segmented on these areas. An alternative scheme to handle texture will be proposed in the next chapter.

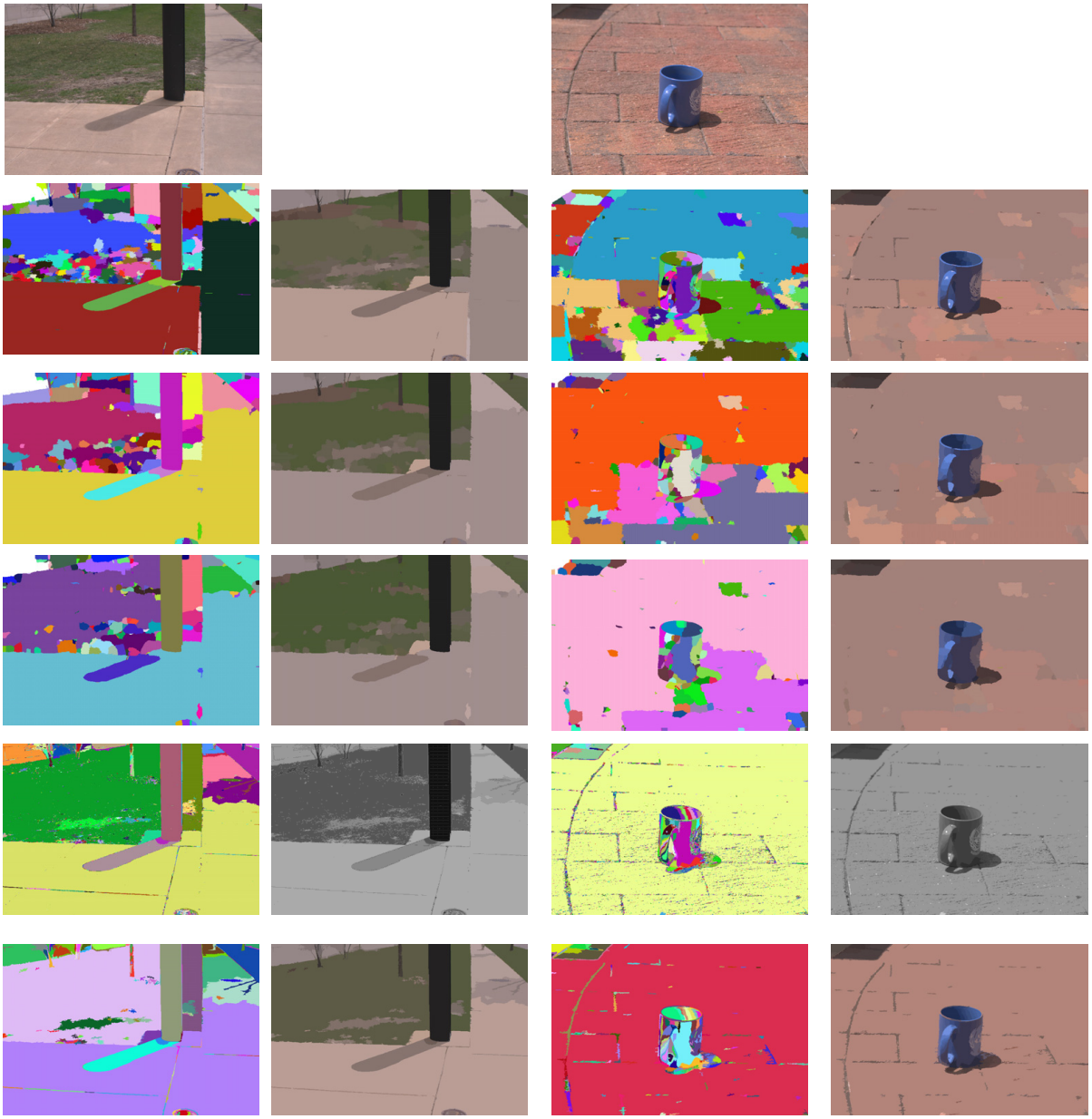


Fig. 4.13. Comparison of the proposed approach with the EDISON system. First row: original RGB image. Second-to-fourth rows: results of the EDISON system with $h_\Omega = 5$,$h_\Omega = 10$ and $h_\Omega = 20$. Fifth and sixth rows: results of the proposed approach before and after post-processing. Odd columns represent the labels with random colours, even columns contain either the RGB-colour medians or the luminance modes for the fifth row. Results in the examples show that the proposed algorithm is able to represent the image in a better or equal way than EDISON operating only on the luminance data. See how fine details as the tree shadow are well conserved by our algorithm while homogeneous areas are also properly shaped.

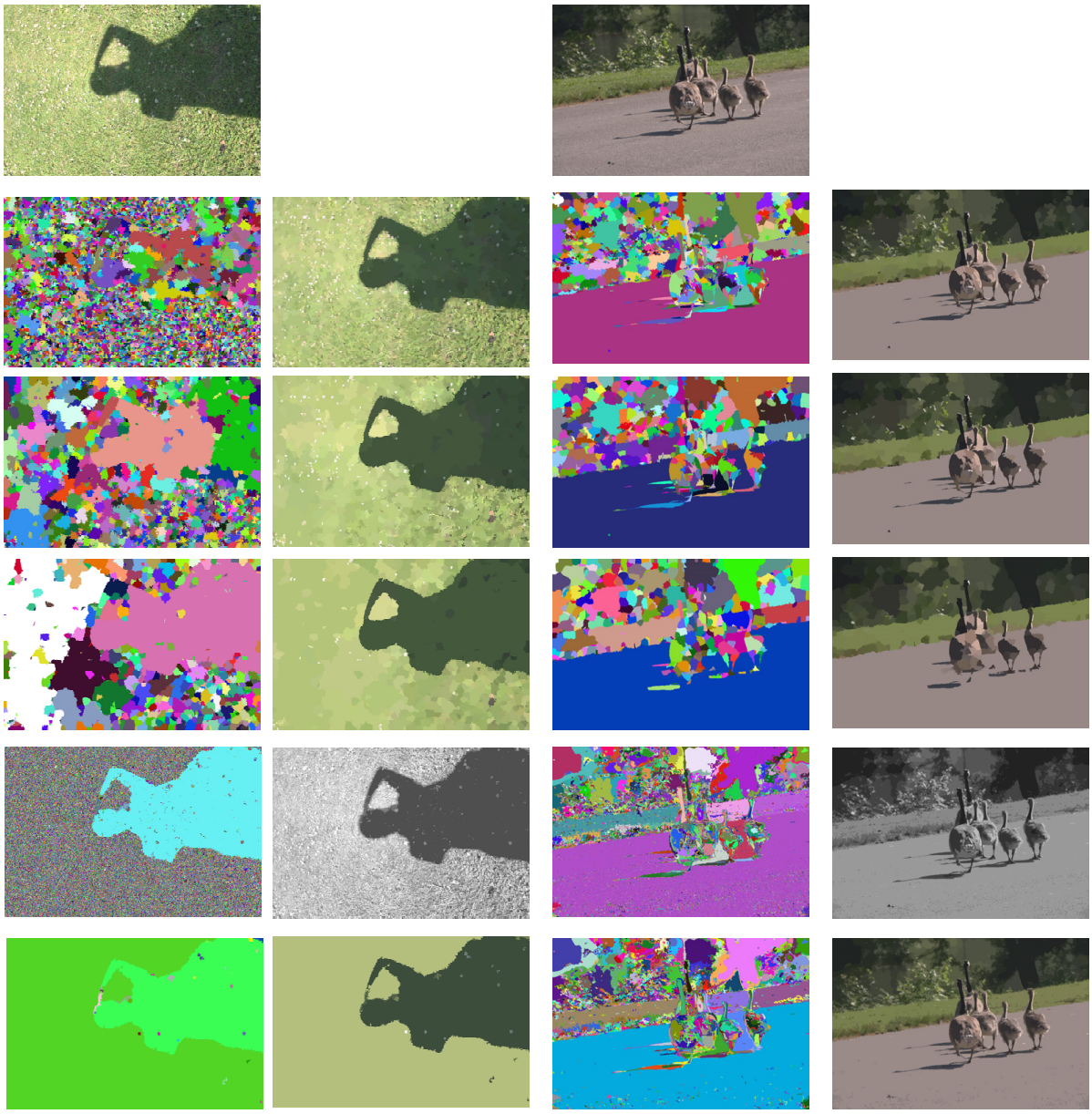


Fig. 4.14. Comparison of the proposed approach with the EDISON system. First row: original RGB image. Second-to-fourth rows: results of the EDISON system with $h_\Omega = 5$,$h_\Omega = 10$ and $h_\Omega = 20$. Fifth and sixth rows: results of the proposed approach before and after post-processing. Odd columns represent the labels with random colours, even columns contain either the RGB-colour medians or the luminance modes for the fifth row. See how the texture-refinement mechanism is able to handle complex textures as the grass in the first column on which the proposed MS-RS approach fails .

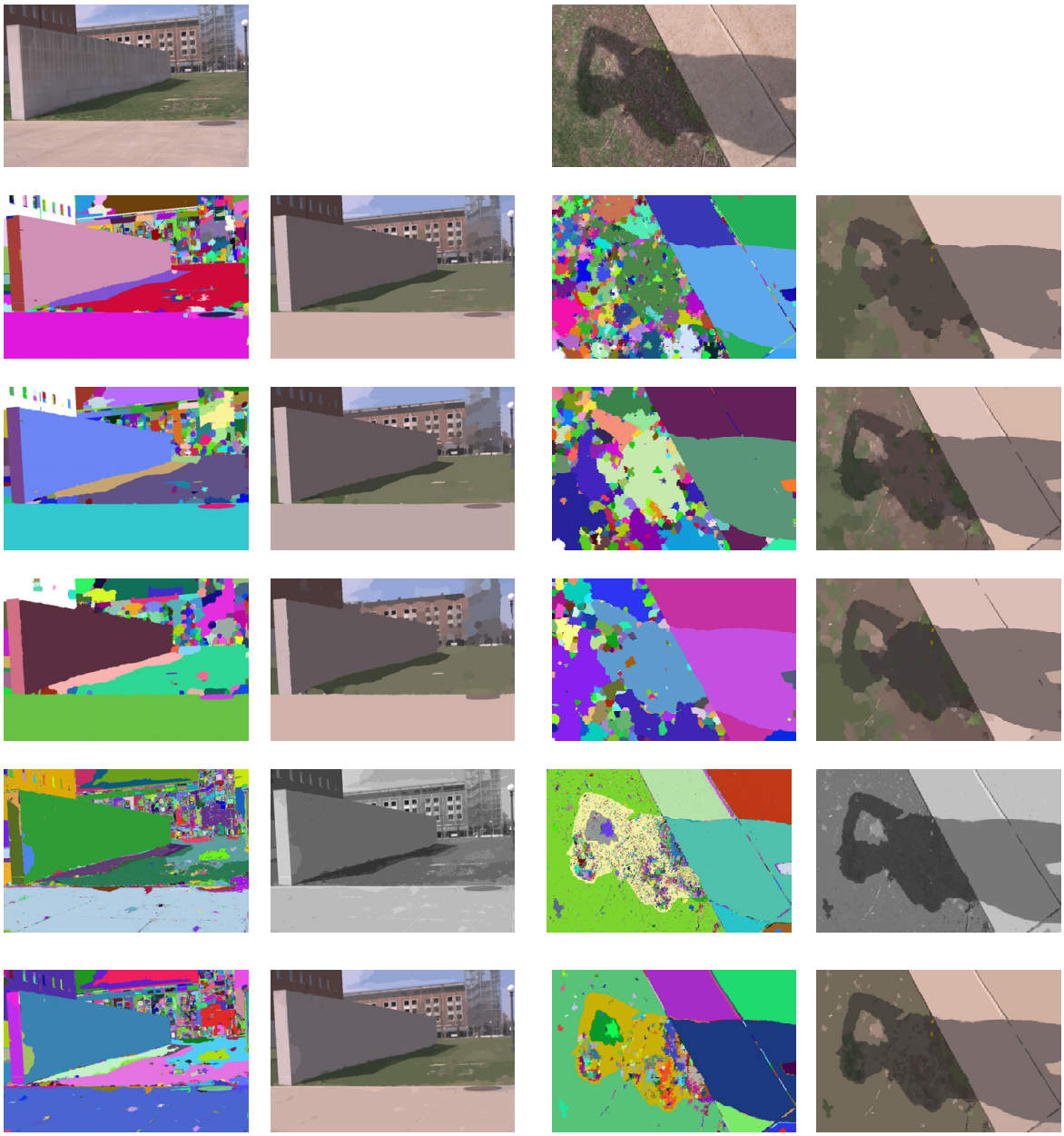


Fig. 4.15. Comparison of the proposed approach with the EDISON system. First row: original RGB image. Second-to-fourth rows: results of the EDISON system with $h_\Omega = 5$,$h_\Omega = 10$ and $h_\Omega = 20$. Fifth and sixth rows: results of the proposed approach before and after post-processing. Odd columns represent the labels with random colours, even columns contain either the RGB-colour medians or the luminance modes for the fifth row. See how the texture-refinement mechanism is again able to handle complex textures as the grass in the third column.

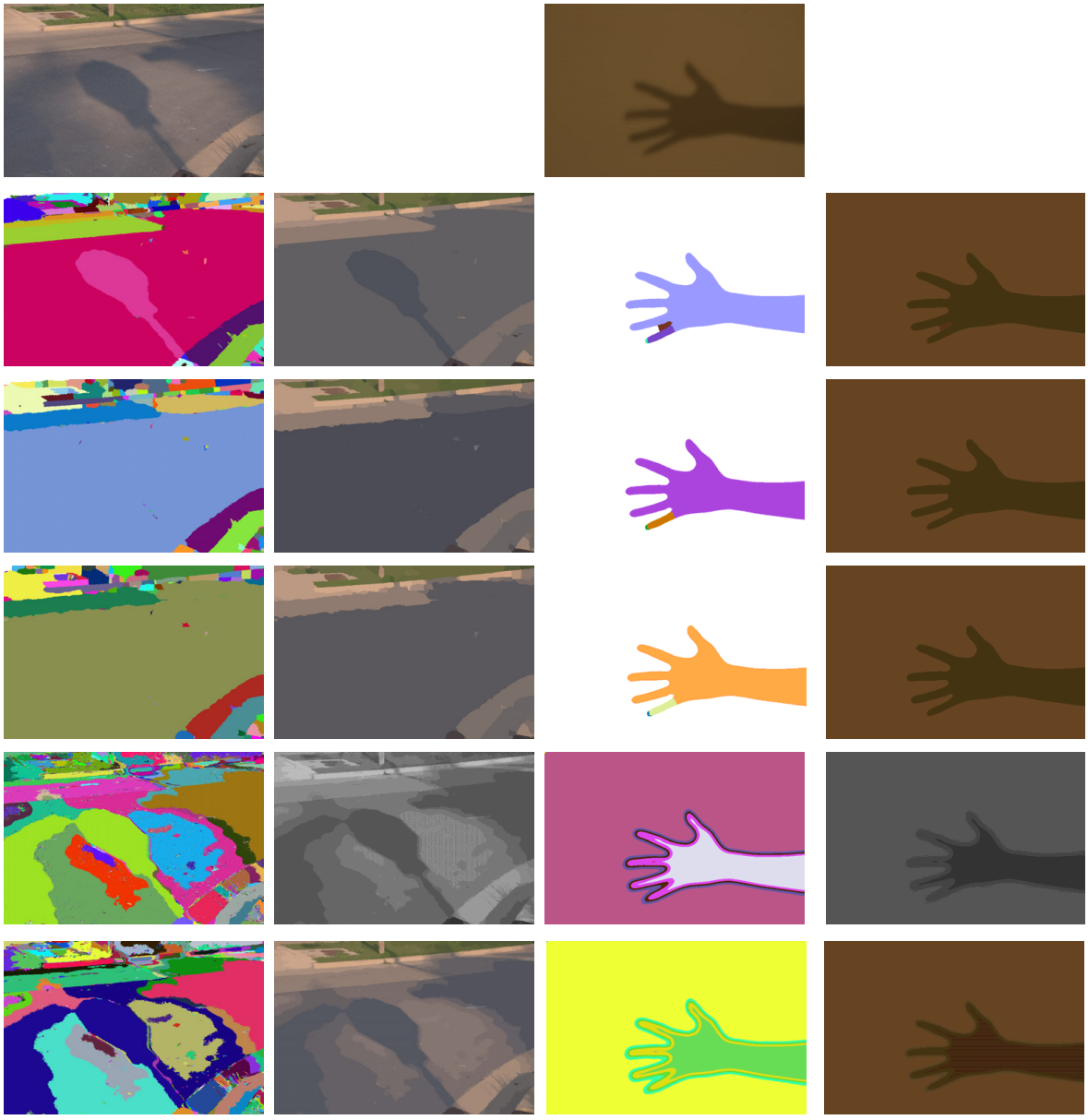


Fig. 4.16. Comparison of the proposed approach with the EDISON system. First row: original RGB image. Second-to-fourth rows: results of the EDISON system with $h_\Omega = 5$,$h_\Omega = 10$ and $h_\Omega = 20$. Fifth and sixth rows: results of the proposed approach before and after post-processing. Odd columns represent the labels with random colours, even columns contain either the RGB-colour medians or the luminance modes for the fifth row. See how fine detail is conserved in results for the proposed method. See also how the proposed method is highly sensible to illumination changes which may be a drawback for some applications but beneficial for some others, e.g. on its use for shadow detection.

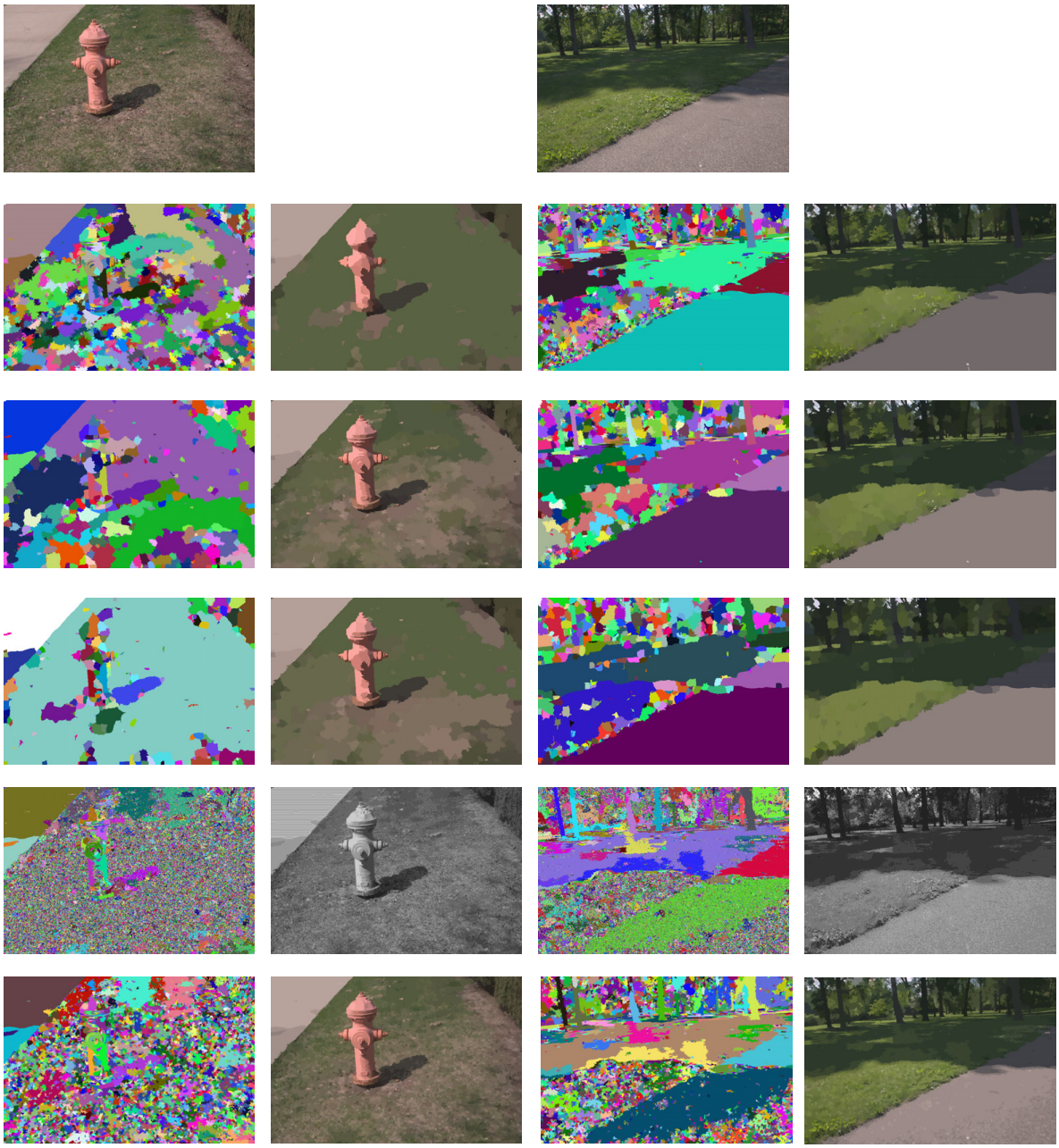


Fig. 4.17. Comparison of the proposed approach with the EDISON system (failure cases). First row: original RGB image. Second-to-fourth rows: results of the EDISON system with $h_\Omega = 5$,$h_\Omega = 10$ and $h_\Omega = 20$. Fifth and sixth rows: results of the proposed approach before and after post-processing. Odd columns represent the labels with random colours, even columns contain either the RGB-colour medians or the luminance modes for the fifth row. Results in the examples show that the proposed algorithm over-segments the image in highly textured areas—even after colour-based post-processing—.

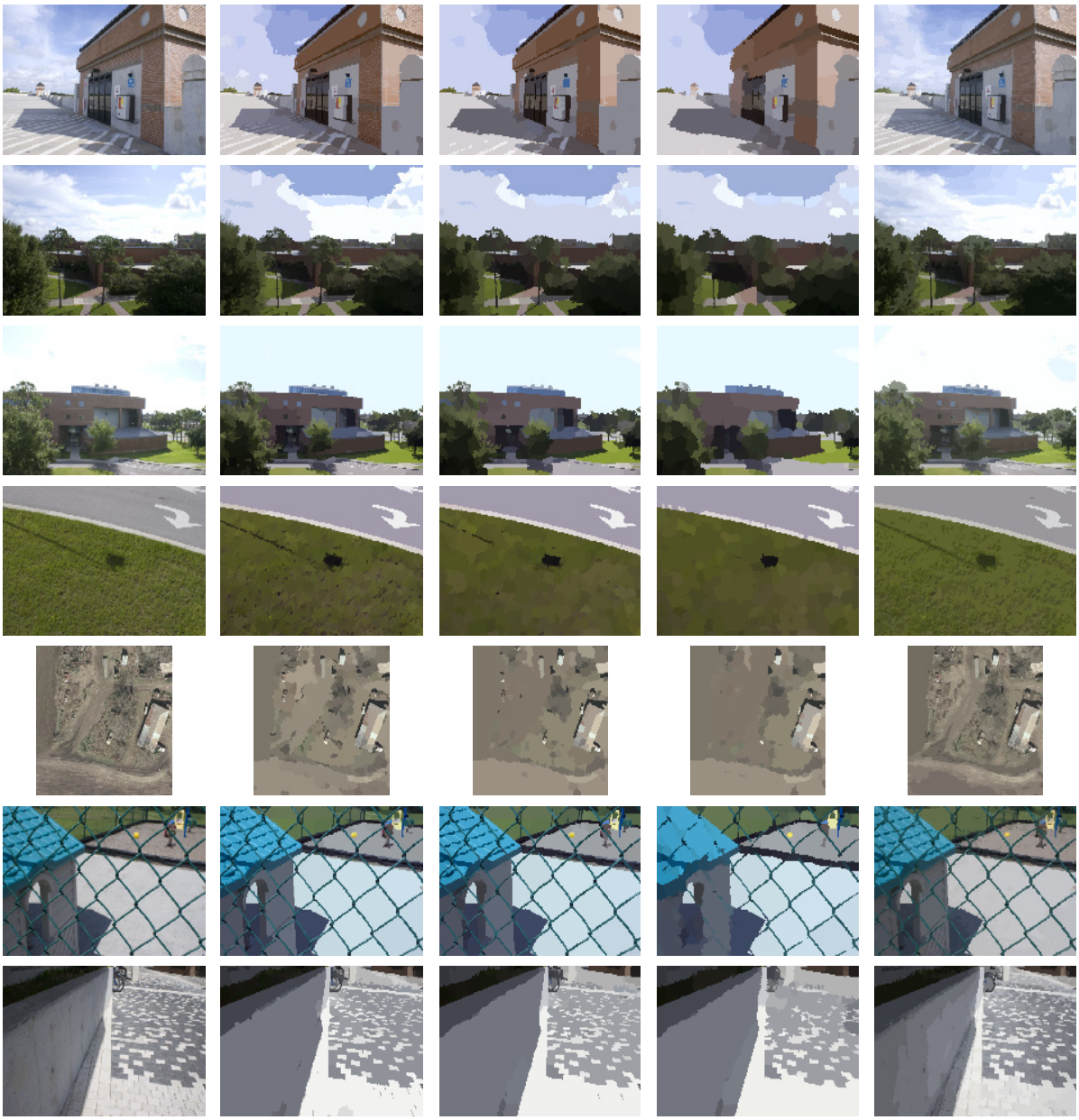


Fig. 4.18. Comparison of the proposed approach with the EDISON system when handling the fine detail in the images. First column: original RGB image. Second-to-fourth columns: results of the EDISON system with $h_\Omega = 5$,$h_\Omega = 10$ and $h_\Omega = 20$. Fifth column: results of the proposed approach after post-processing. Segmentations are represented by the regions RGB-colour median. Results in the examples show that the proposed algorithm is able to respect the fine details better than the EDISON system—even better than EDISON with $h_\Omega = 5$. See examples in the bricks, the sky, the windows, the shadow, the edifications, the fence and the pavement.

Fig. 4.19. Failure cases of the texture refinement method. Right column. Original RGB image. Left column. RGB-colour medians of the proposed approach after post-processing. See how fine details—text in all the examples—are wrongly merged with adjacent surrounding regions.

Chapter 5

Local-variability modelling via the Discrete Cosine Transform

In chapter 4 we have described a RS approach (MS-RS) which was able to conserve the fine detail of the image and most of the scene contours (high recall) but was prone to convey over-segmented partitions (low precision). We conclude that the origin of this problem was the inability of the method to properly handle textured areas.

In this chapter we present a scheme based on the two dimensional Discrete Cosine Transform (from now on, DCT) to model texture in natural images. The DCT is here used as a filter-bank—a family of spatial filters or basis-functions—for discriminative-based texture modelling. Amongst the basis-functions, only a subset are required to represent the majority of the image content for a given pixel. One of the contributions of the chapter is the proposal of a naive method to automatically select this subset of basis-functions for each pixel. This subset is composed of what we name the *relevant* basis-functions. In the proposed texture modelling scheme, each pixel is characterised with the responses of the *relevant* basis-functions on its spatial position. A scheme to compare adjacent pixels is defined in order to detect texture continuity. The *relevant* basis-functions may be different for adjacent pixels; hence, a new metric to compare any two sets of basis-functions responses is required. The definition of this metric constitutes the second main contribution of this chapter. Finally, we propose to derive a contour map based on the selection scheme and the metric. In the map, each pixel is represented by its likelihood of being part of a texture-transition. This constitutes the third and last contribution of this chapter.

The rest of the chapter is organised as follows. We first review existing methods to model local-variability in section 5.1. Then, we introduce the DCT for local-variability modelling in section 5.2. In section 5.3 we describe an scheme to select—for each image pixel—the subset of *relevant* basis-functions. The metric to compare any two DCT set of responses and coefficients is presented in section 5.4. Section 5.5 evaluates the feasibility of establishing a generic selection

of the *relevant* basis-functions with independence of the analysed image. Finally, section 5.6 describes the creation of the contour map and includes some examples to motivate further research in the topic, whereas section 5.7 concludes the chapter.

5.1 Measuring local-variability in natural images

The neighbourhood of a pixel is a representation of the scene surface on which the point projected onto the pixel lies. A RS consisting on the aggregation of adjacent pixels which neighbourhoods are similar may help to identify the image-projection of these scene surfaces. Luminance or colour *constancy* is an indicative of the projection of a flat scene surface without reflectance transitions. However, textured surfaces—those identified by a particular reflectance pattern—may be over-partitioned for a RS method which solely relies on these features.

Local-variability modelling—or local feature measurement—is one of the two main components of texture modelling—the other being the statistical modelling of these measurements—(Xu et al. [2012]). There is a significant amount of studies that provide solutions to model local-variability (Tomita and Tsuji [2013]).

Gradient based approaches—as those proposed in Tsai and Chiu [2008] and Li et al. [2009]—present the limitation that only highly contrasted reflectance patterns are found when a particular threshold is applied on the gradient. Instead, the LBP operator (Ojala et al. [2002]; Heikkilä and Pietikäinen [2006]), considers differences between a given pixel and every other pixel on a predefined neighbourhood. Hence, the LBP can be used to describe the spatial orientation of all of the surrounding edges but, due to data binarisation, ignores absolute edge intensity. Alternatively, local-variability can be described by identifying the spatial filter—from a family of filters—that yields maximum image response on each pixel. This scheme ignores the responses of all the other filters—a relatively recent application of this method is described in Benedek and Szirányi [2008]—. Building on this idea, the known as discriminative Texton-based methods start from the application of a set of filters; each measuring the response of a pixel neighbourhood to a particular spatial pattern. Then, pixels in the same neighbourhood are detected by aggregating responses of all the filters.

According to the recent approaches in texture modelling, Texton-based methods appear to be the current research trend. Let us deepen into these methods.

Textons

Textons can be understood as fundamental micro-structures in natural images. These are known as texture primitives. To remark their relevance, these primitives have been considered the atoms of pre-attentive human visual perception (Julesz [1981]).

We can roughly divide Texton-based methods into generative—which, in our opinion represent the basis of recent developments in Convolutional Neural Networks (CNN)—and discriminative—which are preferred for RS approaches—.

In both schemes, an image $\mathbf{I}$—or an image patch—is described as a linear superposition of weighted and geometrically transformed spatial filters $\mathbf{\Psi} = \{\psi_j,\, j = 1, ..., L_{\mathbf{\Psi}}\}$:

$$\mathbf{I} = \sum_{j=1}^{L_{\mathbf{\Psi}}} c_j \mathcal{T}(\psi_j) \tag{5.1}$$

, where c_j quantifies the contribution associated to $\mathcal{T}(\psi_j)$, and $\mathcal{T}(\psi_j)$ is a geometrical transformation of ψ_j.

The Textons are these weighted and transformed spatial filters $\{c_j \mathcal{T}(\psi_j),\, j = 1, ..., L_{\mathbf{\Psi}}\}$.

The weighting coefficients c_j can be equal to zero, indicating that the information provided by the j^{th} transformed filter does not contribute to the image content.

Operating on the set of filters—or filter-bank—, $\mathbf{\Psi}$, generative methods (Zhu et al. [2005]) aim to obtain the weighting coefficients and the transformation functions $\mathcal{T}(\cdot)$ from the image content. This can be achieved by treating them as latent (hidden) variables and by inferring these variables probabilistically from the image content.

On the contrary, discriminative methods (Malik et al. [2001]; Martin et al. [2004]), characterise each image pixel by a vector of dimensionality $L_{\mathbf{\Psi}}$. Each position of the vector, c_j, contains the response on the pixel to each of the spatial filters. These responses are then clustered to search for common response patterns, implicitly obtaining the Textons as the cluster centres.

The method proposed in this chapter is strongly linked with discriminative methods and it is mainly inspired by them. From here in advance we focus only on this vein of research.

Textons by discriminative methods

Texton-based discriminative (TBD) methods can be applied locally—under a *sparse* approximation—or *densely*.

Sparse approximations consider only the responses of certain pixels in the image to obtain the Textons. These pixels are generally selected under a significance criterion on the scale-space (Lazebnik et al. [2005]; Xu et al. [2012]). S*parse* approximations present the potential benefit of a predetermination of the scale, which allows to define *ad hoc* geometric transformation functions of the spatial filters in agreement with the detected scale. For instance, the transformation can consist in the adaptation, according to the scale parameter, of the spatial area covered by the spatial filters. However, *sparse* methods may be blind to the relevant texture primitives which are representative of the not-analysed pixels.

Dense approximations consider the response of every pixel in the image to obtain the Textons (Leung and Malik [2001]; Malik et al. [2001]; Chantler et al. [2002]). *Dense* approximations are

potentially able to recover all the image texture primitives if a suitable range of spatial filters to cover the whole representation space is defined. Hence, these are usually preferred over *sparse* approximations.

The operation of both schemes is dependent on the design of the basis-functions. In particular, three questions arise:

1. What is the suitable number and nature of the spatial filters to properly represent the texture primitives in an image?
2. How is scale information considered in *dense* approximations?
3. How are responses to these basis-functions combined?

Several solutions to answer these questions have been proposed.

Defining the filters.

On one hand, regarding filter number and nature, several filter-banks have been proposed. The Leung-Malik filter-bank (Leung and Malik [2001]) is composed of 48 filters. It consists of first and second derivatives of Gaussians at 6 orientations and 3 scales making a total of 36 derivative Gaussian filters. Additionally, 8 Laplacian of Gaussian (LoG) filters and 4 Gaussians filters are also part of this filter-bank (depicted in the top-left side of Figure 5.1). Whereas, this filter-bank is usually the preferred for TBD methods, a couple of alternatives have been proposed. For instance, the Schmid filter-bank (Chantler et al. [2002]) consists of 13 rotationally invariant filters. The Maximum Response filter-bank (Varma and Zisserman [2002]) is composed of 38 oriented filters but provides rotation invariance by considering only the maximum response across orientations; hence, taking into account only 8 filter responses. In overall, all of these filter-banks are composed of specific filters that respond to the typical (expected) structures in natural images.

Considering scale information.

On the other hand, as discussed in chapter 3, in Ren [2008]; Arbelaez et al. [2011] it is shown that a multi-scale analysis benefits the identification of texture primitives. Multi-scale information can be easily incorporated in the filter-banks previously described by scaling their spatial response or by increasing the standard deviations of the Gaussian filters that compose them.

Combining the filter responses.

Once the filters are defined, these are applied to the image by spatial convolution. So-obtained responses to the filters are clustered—usually by C-Means—in order to identify common responses and hence, co-neighbour pixels. The belonging relationship of each pixel to each of the

so-obtained clusters is measured and, by assigning the pixel to the cluster to which its response resembles the most; an automatic labelling of the input image into C texture-similar areas is derived.

The methods described in Malik et al. [2001]; Martin et al. [2004]; Arbelaez et al. [2009] are successful examples of this operation path. In these studies, authors propose to cluster the pixels response vectors obtained through the Leung-Malik filter bank into $C = 32$ clusters. The value of C is set as a suitable parameter to cover the local-variability patterns for the majority of natural images. Once the Textons have been shaped, and using the resulting labels as spectral features, texture contours are detected by comparing histograms of these labels on both sides of an hypothetical contour. A graphical flowchart of this method, up to the C-Means stage, is included in Figure 5.1.

Whereas this process is well-founded and has been proven to derive excellent results in the detection of contours in natural images, there are three main issues that remain unclosed.

First, the algorithm would probably benefit for the use of a different C for different images. Second, only some of the filters responses contain significant image information whereas the responses of the others do not represent the image content (flat responses) or are responding to image noise—see the filter responses image in Figure 5.1—. However, the response of all the filters is equally considered in the clustering. Third, the comparison of cluster labels as if these were spectral features is problematic. If each cluster is a Texton, and each pixel is identified with a cluster label , what is the distance between pixel in different clusters,? what is the distance between two Textons and, what is the distance between two spatial filters?

This chapter aims to provide responses to these questions.

Contributions to discriminative-based modelling of local-variability.

The approach that we present in this chapter aims to provide:

1. A scheme for the selection of the *relevant* filters in a filter-bank according to the representativeness of their responses (section 5.3).

2. A metric to compare any two spatial filters and their responses (section 5.4).

3. A scheme to integrate these solutions for contour detection (section 5.6).

These contributions are presented in the scope of a well-established filter-bank: the DCT. The DCT is here used as a filter-bank for TBD modelling. As aforementioned, TBD methods usually rely on filter-banks composed of a set of predefined filters, with each filter designed *ad hoc* for a typical spatial structure in natural scenes. Designing these filters is a key stage in TBD as the overall behaviour of the method depends on it. Four main parameters configure these filters:

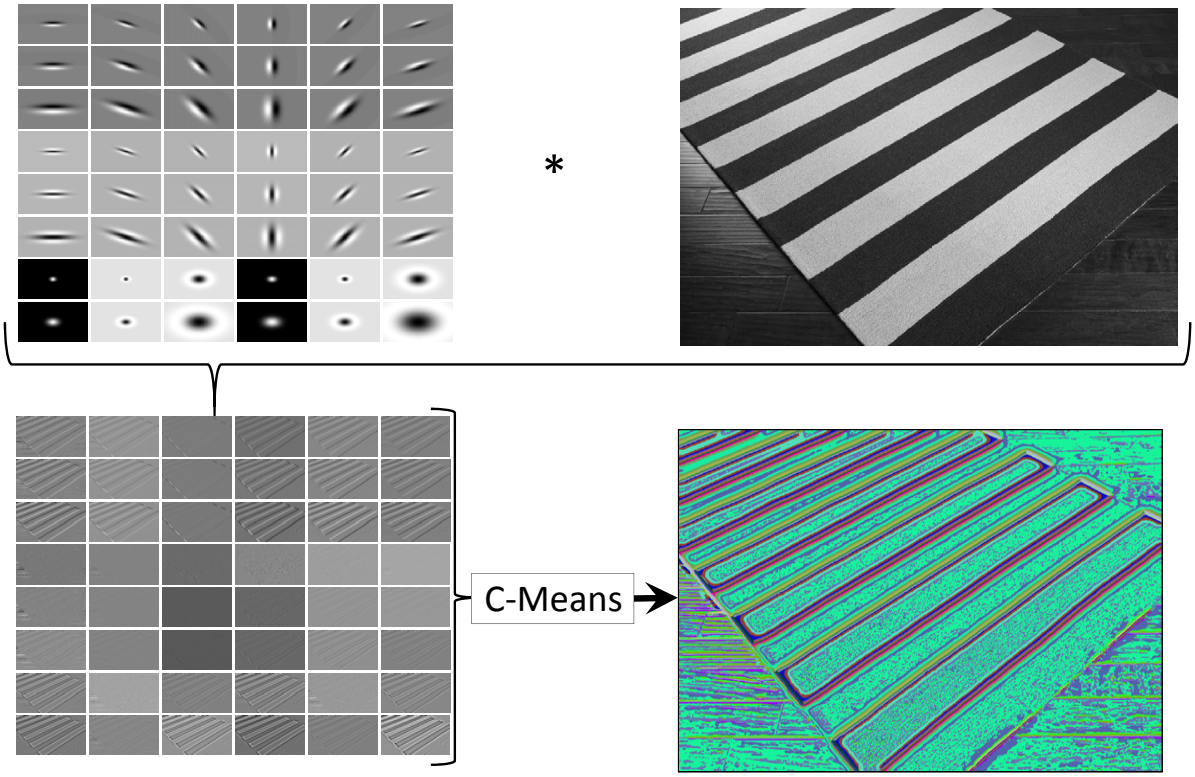


Fig. 5.1. From left to right: (top row) Leung-Malik filter bank. Luminance input image. (bottom row) Pixels responses to each of the filters. C-Means segmentation, with $C = 32$ as suggested in Malik et al. [2001]; Martin et al. [2004]; Arbelaez et al. [2009]. Clusters are represented by random colours. Note how among the responses just some contain relevant information. Flat areas are assigned to a common cluster—represented here by the turquoise label—but textured areas (the wood floor) cannot be grouped into a single cluster as they present texture patterns which might be representative at higher scales.

spatial extend, spatial deviation, orientation and scale. In contrast, using the DCT as a filter-bank reduces the number of parameters to one, the size of the transformed block, which fully defines the rest of the parameters for all the spatial filters in the filter-bank. In the context of the DCT, we call these spatial filters the basis-functions and their responses the coefficients.

In spite of the singularity-blindness of the DCT—it does not respond to the typical structures in natural images—, in this chapter we show that it constitutes a suitable tool to model local-variability.

5.2 The DCT for the representation of local-variability.

In this section we first review the DCT and then briefly enumerate some of its previous uses for local-variability modelling.

The DCT

As generally known, the DCT coefficients, $c(x, y)$ of a WxW pixels square block *centred* at a pixel (u, v) of a scalar image $\mathbf{I}(u, v)$, are computed as in equation 5.2, where $\alpha(0) = \sqrt{\frac{1}{W}}$, $\alpha(x, y \neq 0) = \sqrt{\frac{2}{W}}$, and $0 \leq u, v < W$.

$$c(x,y) = \alpha(x)\alpha(y) \sum_{u=u_0-\frac{W}{2}}^{u_0+\frac{W}{2}-1} \sum_{v=v_0-\frac{W}{2}}^{v_0+\frac{W}{2}-1} \mathbf{I}(u,v) cos\left[\frac{\pi(2u+1)x}{2W}\right] cos\left[\frac{\pi(2v+1)y}{2W}\right] \tag{5.2}$$

, with each DCT coefficient, $c(x, y)$ representing the response of the square block to a basis-function, $\psi_{x,y}$—depicted in Figure 5.2 a) for $W = 8$—.

The DCT has several properties that make it a suitable tool for estimating the local-variability of a pixel (u, v) neighbourhood.

- Each DCT coefficient conveys a measure of the similarity between the $\mathbf{I}(u, v)$ values distribution inside a block *centred* at (u, v) and a directional response determined by the 2D basis-function. The whole block can be seen as a weighted combination of these responses.
- It leads to a set of low-correlated coefficients which are suitable to be modelled independently.
- Illumination changes that have effect on the whole block and are not so strong to occlude variability inside the block, mainly affect the DC coefficient, $c(0, 0)$. Then, a technique not considering this coefficient would be less sensitive to these changes.
- The transform is separable, symmetric and orthogonal. Separability means that each coefficient $c(x, y)$ can be computed in two steps by successive 1-dimensional operations on rows and columns of a block. Symmetry means that row and column operations are functionally identical. Finally, orthogonality stands for such transformations where the inverse transformation matrix is equal to its transpose. All together, these properties allow a fast and efficient computation of the DCT.

From here in advance, we would rely on a dense extraction of the DCT. In particular, we compute the DCT on each image pixel in order to derive a pixel-wise local-variability description. Note that, differently, video and image codecs generally apply the DCT on WxW non overlapping blocks of each frame/image, leading to a WxW response vector per image block. On the contrary,

the extraction of the DCT centred at each pixel (u, v) leads to a WxW vector containing the DCT coefficients: $\mathbf{c}(u, v) = (c(x, y), 0 \leq x, y \leq W)$ for each pixel. Each coefficient is associated to a basis-function $\mathbf{\Psi}(u, v) = (\psi_{x,y}, 0 \leq x, y \leq W)$, which number, nature and size is fully defined by W.

The DCT to measure local-variability

The DCT is a well known operation that has been already used for local-variability modelling and texture indexing. For instance, in Randen and Husoy [1999]; Drimbarean and Whelan [2001], the authors empirically evaluate how descriptors extracted from the DCT better index textured images when compared with descriptors based on Gabor Filters or co-occurrence matrices. The DCT has been also compared with the LBP in terms of its performance on different tasks: as texture descriptors (Paclik et al. [2002]), as appearance models for face detection and recognition (Mendez-Vazquez et al. [2008]), or as tools to code amino acids (Nanni and Lumini [2010]). In these studies, the use of the DCT is motivated by its benefits in information compacting, in description stability and by its ability to condense illumination influence in its low frequency responses.

5.3 Selecting relevant coefficients of the DCT.

In this section we propose a method to discard non-representative basis-functions of a block DCT. We start from the assumption that just a few of the AC coefficients of a block DCT are enough to represent the majority of the relevant information of an image block. This is a common assumption when using the DCT. We name the basis-functions associated with these coefficients: the *relevant* basis-functions.

The method

The proposed method to select the *relevant* basis-functions from a DCT is composed of three stages:

1. **Ordering AC coefficients.** DCT coefficients of each image pixel $\mathbf{c}(u, v)$ are ordered in descending order according to their representativeness. Through this process a set of ordered coefficients for each pixel is obtained. These responses are associated to a set of basis-functions which are ordered in consonance with the responses. Two ordering schemes are explored: *zig-zag* and *ranked.*

2. **Measuring the goodness of reconstruction**. We propose to reconstruct the image at each pixel by considering the N first coefficients in each ordered set of coefficients. This conveys a $N-$partial reconstruction of the image content.

For instance, for the N first coefficients and the DCT transform of the WxW block around the pixel (u, v), the $N-$partially reconstructed WxW square block centred at pixel (u, v), $\tilde{\mathbf{B}}_{(u,v,N)}$, is obtained via the synthesis equation:

$$\tilde{\mathbf{B}}_{(u,v,N)} = \sum_{j=1}^{N} c_{(j)}\psi_{(j)} \tag{5.3}$$

, being $c_{(j)}$ the j^{th} ordered coefficient in the ordered set and $\psi_{(j)}$ its associated basis-function.

Note the similarity of this equation with equation 5.1. In this case, no geometrical transformation $\mathcal{T}(\cdot)$ is applied.

The $N-$partial reconstructed value for the pixel (u, v) can be obtained from $\tilde{\mathbf{B}}_{(u,v)}$ by:

$$\tilde{\mathbf{I}}_N(u, v) = \tilde{\mathbf{B}}_{(u,v,N)}\left(\left\lceil\frac{W}{2}\right\rceil, \left\lceil\frac{W}{2}\right\rceil\right) \tag{5.4}$$

, where $\lceil x\rceil$ stands for the closest integer bigger than x. Note that, in order to ensure that the block has a centre, from here in advance, we constrain W to be an odd number.

The contribution of each ordered DCT coefficient to the image content, $\mathbf{c}(u, v)$, for a particular pixel (u, v), is evaluated by measuring the error (or the similarity) between the original content and each $N-$partial reconstruction of the image. Two comparison measures are explored: Mean Squared Error (MSE) and the Structural SIMilarity Index (SSIM).

3. **Selecting the *relevant* basis-functions**. We aim to obtain the value N^* from which the contribution of the rest of the coefficients in the ordered set is negligible or inconsequential. This N^* is the number of *relevant* basis-functions and is used to truncate the ordered set; hence selecting the subset of *relevant* basis-functions.

Next paragraphs describe these three stages in detail.

Ordering AC coefficients

Two ordering schemes of the AC coefficients are explored. One is an example of the classical scheme followed by the first widely used image and video codecs that rely on the use of the DCT for data compression (*zig-zag*); the other is a simple organisation according to the intensity of the AC coefficients (*ranked*).

The *zig-zag* ordering. In intra-frame encoding of non-interlaced video-sequences, the AC coefficients are sometimes ordered (Le Gall [1991]) according to their potential *relevance* by some sort of *zig-zag* scheme of the basis-functions. For instance, the arrangement:

$$\mathbf{\Psi}_{zz}(u, v) = (\psi_{0,0}, \psi_{0,1}, \psi_{1,0}, \psi_{2,0}, \psi_{1,1}, ..., \psi_{W,W}) \tag{5.5}$$

, conveys an ordering:

$$\mathbf{c}_{zz}(u,v) = (c(0,0), c(0,1), c(1,0), c(2,0), c(1,1), ..., c(W,W)). \tag{5.6}$$

A so-defined arrangement does not require any prior analysis on the image content and, hence, can be straightly standardised. However, the first coefficients in $\mathbf{c}_{zz}(u,v)$ are not necessarily those that represent the majority of the information in the DCT (as we experimentally prove in section 5.5).

The *ranked* ordering. In order to quantitatively measure the representativeness of every AC coefficient, let us define the *relevance*, $w(x,y)$, of an AC coefficient within the scope of its DCT block as a function of its relative *intensity* respect to the total transformed *intensity* in the block, i.e.:

$$w(x,y) = \frac{|c(x,y)|}{\sum_{x=0}^{W}\sum_{y=0}^{W}|c(x,y)|} \tag{5.7}$$

Being $c_{(WxW-N+1)}$ the N^{th} order statistics, the N most *intense* AC coefficients of the DCT transform of a WxW square block centred at pixel (u,v)—or the $N-$top *ranked* AC coefficients—can be obtained by:

$$\mathbf{c}_{ranked}(u,v,W,N) = (c_{(WxW)}, c_{(WxW-1)}, ..., c_{(WxW-N+1)}) \tag{5.8}$$

, which associated DCT basis-functions are arranged according to $\mathbf{c}_{ranked,N}(u,v,W,N)$ as:

$$\mathbf{\Psi}_{ranked}(u,v,W,N) = (\psi_{(WxW)}, \psi_{(WxW-1)}, ..., \psi_{(WxW-N+1)}) \tag{5.9}$$

, with $\psi_{(j)} = \psi_{x_0,y_0}$ being the basis-function to which the coefficient $c_{(j)} = c(x_0,y_0)$ represents the response and being $w_{(j)} = w(x_0,y_0)$ the j^{th} highest relative *intensity* in the *relevance* set: $\mathbf{w}(u,v) = (w(x,y),\ 0 \leq x,y \leq W)$.

In section 5.5 we evaluate the behaviour of these two ordering schemes on a set of training images.

Measuring the goodness of reconstruction.

We aim to use a fidelity signal through which compare two images: the original image $\mathbf{I}$ and each partially reconstructed image $\tilde{\mathbf{I}}_N$. We require the fidelity signal to represent a quantitative score to describe the degree of similarity or, conversely, the level of distortion between the two images.

The MSE has been the preferred comparison signal for this purpose; the MSE of each $N-$partial reconstruction, $MSE(N)$, can be defined as:

$$MSE(N) = \frac{1}{UxV} \sum_{u=1}^{U} \sum_{v=1}^{V} (\mathbf{I}(u,v) - \tilde{\mathbf{I}}_N(u,v))^2 \tag{5.10}$$

, where UxV is the image resolution.

In Wang and Bovik [2009], the problems of MSE when comparing images are discussed and the use of alternative comparison signals is strongly motivated. Among these alternatives, the SSIM (Wang et al. [2004]) is claimed to overcome most of the inaccuracies of a squared-error comparison and to provide a better measure of the images inter-similarity.

For a pixel (u,v) the SSIM index comparing $\mathbf{I}(u,v)$ and $\tilde{\mathbf{I}}_N(u,v)$ is extracted in patches around (u,v) on each image: $\mathbf{b}$ and $\tilde{\mathbf{b}}$. The SSIM value for such pixel. $SSIM(u,v)$, can be computed—under some assumptions—as:

$$SSIM(u,v) = \frac{(2\mu_{\mathbf{b}}\mu_{\tilde{\mathbf{b}}} + C_1)\left(2\sigma_{\mathbf{b},\tilde{\mathbf{b}}} + C_2\right)}{\left(\mu_{\mathbf{b}}^2 + \mu_{\tilde{\mathbf{b}}}^2 + C_2\right)\left(\sigma_{\mathbf{b}}^2 + \sigma_{\tilde{\mathbf{b}}}^2 + C_2\right)} \tag{5.11}$$

, where $\mu_{\mathbf{b}}$ and $\sigma_{\mathbf{b}}$ are the local sample mean and standard deviation of patch $\mathbf{b}$, $\sigma_{\mathbf{b},\tilde{\mathbf{b}}}$ is the cross-correlation between $\mathbf{b}$ and $\tilde{\mathbf{b}}$, $C_1 = (0.01\varrho)^2$, $C_2 = (0.03\varrho)^2$ and ϱ is the dynamic range of the image, i.e, $\varrho = max(\mathbf{I}) - min(\mathbf{I})$. Additional details about the SSIM index can be consulted in Wang et al. [2004] .

A global similarity index between the entire image and each of its $N-$partial reconstruction is then extracted by averaging the SSIM values:

$$M - SSIM(N) = \frac{1}{UxV} \sum_{u=1}^{U} \sum_{v=1}^{V} SSIM(u,v) \tag{5.12}$$

In section 5.5 we evaluate the behaviour of these two comparison signals on a training set of images.

Selecting the *relevant* basis-functions.

We propose to identify the *relevant* basis-functions by observing the evolution of the fidelity signals $MSE(N)$, and $M - SSIM(N)$. In particular, the aim is to locate N^*, the position in the ordered set of coefficients from which the rest of the coefficients can be discarded.

The location of N^* results in the selection of the N^* first coefficients in one of the ordering schemes. For instance, for the *ranked* ordering, a pixel (u,v) and a DCT filter-bank defined by W, the location of N^* results in the selection of the set of coefficients, $\mathbf{c}_{ranked}(u,v,W,N^*)$, and of the set of associated basis-functions, $\mathbf{\Psi}_{ranked}(u,v,W,N^*)$.

Let us assume that $MSE(N)$ is a non-increasing function and $M - SSIM(N)$ a non-decreasing function of N with independence of the ordering scheme used to obtained them.

Furthermore, let us also assume that $MSE(N)$ [$M - SSIM(N)$] is an L-shaped [inverse L-shaped] function of N, i.e, a function that shows a sharp fall [ascent] after which values remain low [high] for subsequent values of N. In this case, a good selection for N^* will be the elbow or corner of these functions. Both assumptions are experimentally evaluated in section 5.5.

The result of this method is the selection of the set of *relevant* basis-functions for a given DCT transform. In particular, for a predefined set of DCT transforms: $\mathbf{W} = \{W_i\}$ the method conveys a set of *relevant* basis-functions by the identification of the set of positions: $\mathbf{N}^* = \{N_i^*\}$.

Experimentally, in section 5.5 we will show that, if small errors are tolerated, the value of N^* can be selected the same with independence of the pixel and the analysed image. Nevertheless, the process here described can be instead applied locally on each image in order to improve the quality of estimation of the set $\mathbf{N}^*$, at the expense of increasing the computational cost of the whole solution.

A particular value of N^* will select the same set of basis-functions for all the image pixels if these are sorted via the *zig-zag* ordering scheme. Independently of the value of N^* this scheme will allow the straight comparison of the coefficients of any two image pixels (as done in [Ji and Park, 2000; Lamarre and Clark, 2002; Tachizaki et al., 2009]).

However, a particular value of N^* might produce the selection of a different set of basis-functions for different pixels in the image if the *ranked* ordering scheme was used for their arrangement. In this case, the straight comparison of the coefficients would be senseless, as these might be the image responses to different basis-functions. To overcome this problem a metric to compare any two N^*-length sets of coefficients is proposed in the next section.

5.4 DCT-based comparison of pixel-wise local-variability descriptions

In this section we propose a metric to compare coefficients obtained as responses of different basis-functions in the DCT filter-bank. Each AC coefficient, $c\,(x, y)$, represents the response to a basis-function, $\psi_{x,y}$. Hence, independently of the AC coefficient value, dissimilarity evaluation first requires a measure of the similarity between every pair of basis-functions. We here propose a simple estimation of such subjective similarity, attending to spatial variability rhythm and direction, and weighting these in a well-balanced fashion. Then we prove that such similarity measure is in fact a metric. This metric is used to define a new metric that also accounts for coefficient intensity. This last metric allows to compare the sets of representative responses of any two image pixels with the only condition that they have to be of the same length. Finally, we define a third metric, which builds on the two previous, and combines information from a predefine set of DCT transforms, $\mathbf{W} = \{W_i\}$.

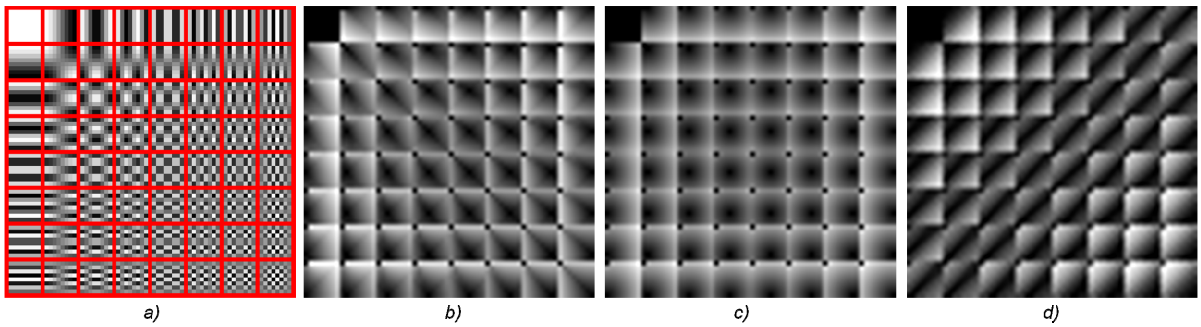


Fig. 5.2. a) Representation of the DCT basis-functions for $W = 8$, b) Metric evaluation between every single AC basis-function and all the other ones, c) 2D Euclidean distance between every single AC basis-function and all the other ones, d) 1D Euclidean distance between every single zigzag ordered AC basis-function and all the other ones.

A metric to compare any two DCT basis-functions

The DCT provides a good, but unbalanced, combination of spatial orientation and intensity of local-variability information. To our knowledge, there are no previous approaches describing pixel-wise local-variability with a set of coefficients resulting from a variable set of basis-functions of a DCT filter-bank. In our opinion, this might be due to the problematic involved in establishing relations between the basis-functions.

Analysing equation 5.2 and Figure 5.2 a), it can be observed that, by increasing the values of x and y independently—increasing one by setting the other to a fixed value—, the resulting basis-functions increase in terms of spatial variability rhythm for a set direction. Alternatively, changing the values of x and y at the same time also results in a change in the basis-function direction.

Considering the classical 2D representation of the DCT basis-functions—see Figure 5.2 a)—, we obtain a measure of the *distance* between two of these basis-functions following equation 5.13, where $a \vee b$ stands for the maximum of a and b, and $atan(a)$ stands for the arc tangent of a.

$$M\left[\psi_{x_1,y_1}, \psi_{x_2,y_2}\right] = k_1\left[\mid x_1 - x_2 \mid \vee \mid y_1 - y_2 \mid\right] + k_2\left[\mid atan\left(\frac{x_1}{y_1}\right) - atan\left(\frac{x_2}{y_2}\right) \mid\right] \tag{5.13}$$

The non negative weight factors k_1 and k_2 are set to equally weight both terms of the equation, which intends to formalise that patterns with maximum difference in variability orientation—, i.e., orthogonal orientations—are considered as different as those with equal orientation but maximum difference in variability rhythm.

Observing that the first term in equation 5.13 takes values in $[0 : W - 1]$, and the second

one varies in $[0 : \pi/2]$, we can set $k_1 = 1$ and $k_2 = (W-1)\pi/2$, but other combinations of k_1 and k_2 keeping balance between the two parts of the equation would be also valid. The proposed measure fulfils the properties of non-negativity, positive definition, symmetry and sub-additivity—proofs of these properties are included below—; hence, we can call it a metric.

Lemma The similarity measure defined in equation 5.13 where $k_1, k_2 \geq 0$, and $\psi_{x_1,y_1}, \psi_{x_2,y_2}$ have both components positive, is a metric.

Proof.

- **Non-negativity:** trivial as $k_1, k_2 \geq 0$
- **Positive definiteness:** if $c_1 = c_2$ then $M\left[\psi_{x_1,v_1}, \psi_{x_2,y_2}\right] = k_1 0 + k_2 |atan(1) - atan(1)| = 0$. If $M\left[\psi_{x_1,y_1}, \psi_{x_2,y_2}\right] = 0$ then if $k_1 > 0$, in order to have the first part equal to zero we should have $\psi_{x_1,y_1} = \psi_{x_2,y_2}$. If $k_2 > 0$ then we have for the second factor $\psi_{x_1,y_1} = n\psi_{x_2,y_2}$ for every n. That is: the measure is a metric if $k_1, k_2 > 0$. Otherwise, if $k_1 = 0$, and $k_2 > 0$, M is a pseudo-metric.
- **Symmetry:** $M\left[\psi_{x_1,y_1}, \psi_{x_2,y_2}\right] = k_1 \left[\mid x_1 - x_2 \mid \vee \mid y_1 - y_2 \mid\right] + k_2 \left[\mid atan(\frac{x_1}{y_1}) - atan(\frac{x_2}{y_2}) \mid\right] =$ $k_1 \left[\mid x_2 - x_1 \mid \vee \mid y_2 - y_1 \mid\right] + k_2 \left[\mid atan(\frac{x_2}{y_2}) - atan(\frac{x_1}{y_1}) \mid\right] = M\left[\psi_{x_2,y_2}, \psi_{x_1,y_1}\right]$, thanks to the use of the absolute values of the differences.
- **Sub-additivity:** $M\left[\psi_{x_1,y_1}, \psi_{x_2,y_2}\right] = k_1 \left[\mid x_1 - x_2 \mid \vee \mid y_1 - y_2 \mid\right] + k_2 \left[\mid atan(\frac{x_1}{y_1}) - atan(\frac{x_2}{y_2}) \mid\right]$
  $= k_1 \left[\mid x_1 - x_3 + x_3 - x_2 \mid \vee \mid y_1 - y_3 + y_3 - y_2 \mid\right]$
  $+ k_2 \left[\mid atan(\frac{x_1}{y_1}) - atan(\frac{x_3}{y_3}) + atan(\frac{x_3}{y_3}) - atan(\frac{x_2}{y_2}) \mid\right]$
  $\leq k_1 \left[(\mid x_1 - x_3 \mid + \mid x_3 - x_2 \mid) \vee (\mid y_1 - y_3 \mid + \mid y_3 - y_2 \mid)\right]$
  $+ k_2 \left[\mid atan(\frac{x_1}{y_1}) - atan(\frac{x_3}{y_3}) \mid\right] + k_2 \left[\mid atan(\frac{x_3}{y_3}) - atan(\frac{x_2}{y_2}) \mid\right]$
  $\leq k_1 \left[\mid x_1 - x_3 \mid \vee \mid y_1 - y_3 \mid\right] + k_1 \left[\mid x_3 - x_2 \mid \vee \mid y_3 - y_2 \mid\right]$
  $+ k_2 \left[\mid atan(\frac{x_1}{y_1}) - atan(\frac{x_3}{y_3}) \mid\right] + k_2 \left[\mid atan(\frac{x_3}{y_3}) - atan(\frac{x_2}{y_2}) \mid\right]$
  $\leq M\left[\psi_{x_1,y_1}, \psi_{x_3,y_3}\right] + M\left[\psi_{x_3,y_3}, \psi_{x_2,y_2}\right]$

, where we used in the first inequality the triangle inequality for the absolute distance in $\mathbb{R}$ given by $d(x, y) = |x - y|$ and in the second inequality two times the simple fact that: $(|a| \vee (|b| + |c|)) \leqq (|a| \vee |b|) + (|a| \vee |c|)$, for every $a, b, c \in \mathbb{R}$.

The proposed metric can be visually inspected in Figure 5.2 b). The metric evaluated for every pair of AC basis-functions is plotted block-wise, that is, it is organised in a $WxW-$blocks grey-level image. Each block's pixel presents the distance—the higher the brighter—between the co-located basis-function displayed in Figure 5.2 a) and all the other $WxW - 1$ functions—including self-similarity and excluding similarity with the DC basis-function, which is

set to zero or black—. The block corresponding to the DC coefficient is also set to zero as it will be unused by the current method.

In spite of our restriction on odd values for W, and for visualisation purposes, we have set $W = 8$ and scaled the resulting images. Observe, for instance, that $\psi_{0,1}$ results as different from $\psi_{1,0}$—just due to variability direction—as from $\psi_{0,7}$ —just due to variability rhythm—.

An intuitive but in this case senseless alternative is to use the Euclidean distance between the basis-functions positions, i.e. (x, y), in a 2D vector space; this is illustrated in Figure 5.2 c). Observe that in this case $\psi_{0,1}$ results relatively *similar* to $\psi_{1,0}$, while representing orthogonal patterns.

Finally, we also include in Figure 5.2 d) the 1D Euclidean distance between every single AC basis-function and all the other ones, ordered following the *zig-zag* ordering scheme described in previous section. This scheme is used, for instance, in the computation of the Colour Layout descriptor of MPEG-7 [Kasutani and Yamada, 2001]. Again, orthogonal patterns are very close in the distance space—observe the similarity between $\psi_{0,1}$ and $\psi_{1,0}$ which are separated by the minimum distance step—.

A metric to compare any two equal-length sets of coefficients.

Let $\mathbf{c}_{ranked}(u, v, W_i, N_i^*)$ be the set of N_i^* *ranked* orderer set of DCT coefficients of a $W_i x W_i-$block around pixel $\mathbf{p} = (u, v)$ in the shape of equation 5.8 but excluding the DC coefficient $c(0, 0)$ in the ordering. Let $\mathbf{\Psi}_{ranked}(u, v, W_i, N_i^*)$ be the set of associated basis-functions (following equation 5.9).

Let $\mathbf{c'}_{ranked}(u', v', W_i, N_i^*)$ and $\mathbf{\Psi'}_{ranked}(u', v', W_i, N_i^*)$ be the equivalents of these sets for pixel $\mathbf{p'} = (u', v')$.

According to equation 5.13, a suitable metric to compare pixels $\mathbf{p}$ and $\mathbf{p'}$ in terms of the local-variability $(L - V)$ around them can be derived by averaging the distance between the basis-functions that describe their local-variability:

$$d_{L-V}(\mathbf{p}, \mathbf{p'}, W_i) = \frac{1}{N_i^*} \sum_{j=0}^{N_i^*-1} M\left[\psi_{(W_i x W_i - j)}, \psi'_{(W_i x W_i - j)}\right] \tag{5.14}$$

However, such comparison does not account for the numeric responses of the blocks to the basis-functions, i.e. for the scalar value of the coefficients. Whereas the coefficients are intrinsically considered in the ordering stage; a more explicit scheme to introduce them in the comparison is:

$$d_{L-V}(\mathbf{p}, \mathbf{p'}, W_i) = \frac{1}{N_i^*} \sum_{j=0}^{N_i^*-1} \alpha[c_{(W_i x W_i - j)}, c'_{(W_i x W_i - j)}] \cdot M\left[\psi_{(W_i x W_i - j)}, \psi'_{(W_i x W_i - j)}\right] \tag{5.15}$$

, where:

$$\alpha[c_{(j)}, c'_{(j)}] = 1 + \left(max\left(w_{(j)}, w'_{(j)}\right) - min\left(w_{(j)}, w'_{(j)}\right)\right) \tag{5.16}$$

, with $w_{(j)}$ extracted as in equation 5.7.

Note that, the minimum function $min(\cdot)$ has been used to conserve the symmetry property of metrics, i.e. to ensure that: $d_{L-V}(\mathbf{p}, \mathbf{p}', W_i) = d_{L-V}(\mathbf{p}', \mathbf{p}, W_i)$.

Observe that, in the case of $w_{(j)} = w'_{(j)}$, equation 5.16 stands: $\alpha[c_{(j)}, c'_{(j)}] = 1$; hence, the corresponding j^{th} term in equation 5.15 adds $M\left[\psi_{(W_i x W_i - j)}, \psi'_{(W_i x W_i - j)}\right]$ to $d_{L-V}(\mathbf{p}, \mathbf{p}', W_i)$.

A metric to include multi-scale information in the comparison

Images structures can be representative at different scales. In order to account for multi-scale information, we propose to explore different values of W_i to measure local-variability. The value of W_i defines the size of the neighbourhood used for the transformation; hence, W_i controls the scale on which to model local-variability.

We propose to compare local-variability descriptions at seven scales: $W_i \in [3, 15]$, W_i *odd*. The cut-off value N_i^* is a dependent parameter of the block size. In section 5.5 we experimentally compute suitable N_i^* values for each analysed W_i.

Figure 5.3 exemplifies how the scale of the structure determines the comparison. In the middle column of the Figure, we have included the distance $d_{L-V}(\mathbf{p}, \mathbf{p}', \mathrm{W_i})$ between a pixel $\mathbf{p}$—indicated by a red dot in the top-left column of the Figure—and every other image pixel for selected values of W_i. The experiment is carried out for four different images and searching for four different variability patterns—row-wise in the Figure—.

The distance $d_{L-V}(\mathbf{p}, \mathbf{p}', \mathrm{W_i})$ is lower—to different degrees—for pixels which surrounding local-variability resembles that of pixel $\mathbf{p}$. However, this effect is visible just at some of the scales. Let us discuss this visibility in terms of the *isolation* of the local-variability patterns. Let us define *isolation* as the association of these patterns to lower distances than their surrounding pixels.

For instance, for the first row, on which we search for the leopard skin pattern, the distance between $\mathbf{p}$ and the pixels associated to the leopard skin is somehow stable along the showed scales. Nevertheless, the leopard seems to be better *isolated* from the background on high scales. A similar effect is observed for the tree branches in the second image. However, in this case, the branches region is just partially and slightly *isolated* on a couple of scales ($W_i = 7$ and $W_i = 13$). For the tiger, the better *isolation* seems to be achieved for $W_i = 13$. Nevertheless, the tiger silhouette can be identified at almost every scale. However, the tree in the tiger image is also assigned lower distances. The beaver image is probably the one of highest complexity. There are several different variability patterns in the scene, and the beaver fur also present different

variability patterns. Searching for one of these patterns results in lower distances for the pixels in the pattern at some scales—see for instance $W_i = 13$—.

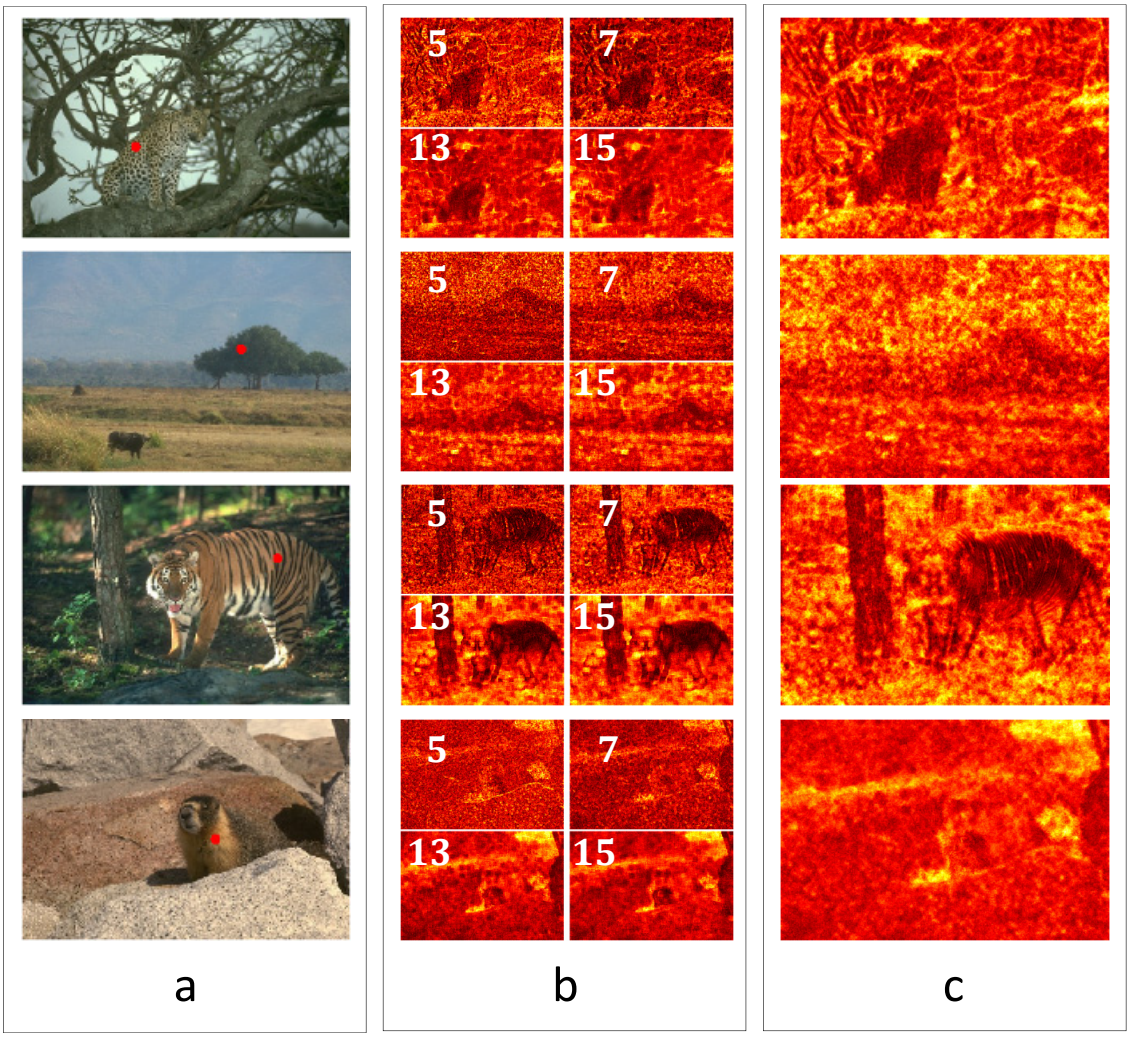


Fig. 5.3. Multi-scale DCT comparison. Column a. Example image with searched pixel indicated by a red dot. Column b. Distance (through equation 5.15) of such pixel and every other pixel in the image (the darker the lower) for different scale values $W_i = [5, 7, 13, 15]$ and associated N_i^* values (see Table 5.1). Column c. Aggregated distance for all the scales (through equation 5.17). See text for discussion.

We propose to disambiguate information from multiple scales by aggregating the distances obtained for all the scales, i.e.:

$$d_{L-V}(\mathbf{p}, \mathbf{p}') = \sum_{i}^{|\mathbf{W}|} \left(d_{L-V}(\mathbf{p}, \mathbf{p}', W_i) \right) \tag{5.17}$$

This constitutes our final local-variability based pixel comparison. Experimentally, we have assessed that this combination results in a better discrimination that alternative schemes, e.g. selecting the minimum distance obtained when analysing all the scales. Incidentally, multi-scale distance aggregation is also the scheme followed by Arbelaez et al. [2011].

Figure 5.3 includes this final distance on its top-right column. Whereas the leopard and the tiger skin patterns appear to be well *isolated*, the tree branches and the beaver skin *isolation* is less evident. However, we will show that a local analysis (comparing a pixel with its neighbouring pixels, not with the whole image) may provide better results than those in Figure 5.3.

5.5 Experimental selection of the *relevant* coefficients and associated basis-functions.

This section addresses the topic of coefficients representativeness; in particular, we aim to experimentally answer the following question: how many coefficients are required to reliably represent a natural image? The advantage of the DCT is that, in general, the reconstruction of the image considering only a few of the coefficients is prone to convey a good approximation to the original image content Guleryuz [2007].

By means of the next experiments we aim to show that the number of *relevant* basis-functions, N_i^*, via which each pixel set of coefficients is truncated, is somehow stable for several images.

Experiments description

We compare the operation of the two ordering schemes proposed in section 5.3: *zig-zag* and *ranked* ordering. These are compared by measuring the goodness of reconstruction of the $N-$partial reconstructions attained by using equation 5.3.

These $N-$partial reconstructions are computed: for every N (given W_i, $N \in [1, W_i x W_i] \subset \mathbb{Z}$); for every pixel (u, v); and using—for each pixel—$c_{(j)}$ and $\psi_{(j)}$ from equations 5.5 and 5.6 for the *zig-zag* ordering scheme; and from equations 5.8 and 5.9 for the *ranked* ordering scheme.

The comparison is performed in terms of the two fidelity signals defined in section 5.3: MSE (equation 5.10) and SSIM (equations 5.11 and 5.12). Hence, a total of four evaluations result for each experiment (one for each pair of ordering scheme and fidelity signal).

Three experiments (**Ex.**) are carried out on the content available in the training set of the BSD500 data-set (Martin et al. [2001]; Arbelaez et al. [2011]):

Ex.1 Consist in the analysis of a single image given a set of basis-functions (defined by W_i). Results for this experiment are included in Figure 5.4. The top part of the Figure (its first

two rows) includes examples of the $N-$partial reconstructions of the image for the two ordering schemes and $W_i = 9$ (for the experiment we choose the mid scale value). The bottom part of the Figure (last row) depicts the MSE and the M-SSIM associated to the ordering schemes as functions of N. The standard deviations of the squared error (SE) and the SSIM for all the image pixels are also included by means of an error-bar graphic—the larger the bar, the higher the deviation—.

Ex.2 Consist in the analysis of the whole set of images in the data-set given a set of basis-functions (defined by W_i). Results for this experiment are depicted in Figures 5.5 and 5.6. Results in Figure 5.5 are mesh representations of the MSE and the M-SSIM measures as functions of N for $W_i = 9$ and the 200 images in the training set. Results in Figure 5.6 represent the image-averaged of these MSE and M-SSIM measures. In the Figure, the length of the bars in the error-bar graphic represents the standard deviation of the SE and the SSIM on all the pixels in the 200 training images. The SE and the SSIM extracted for each image pixel are not, in general, independent variables when extracted for a group of images. Hence, the calculation of the standard deviation cannot be done by aggregating per-image variances—i.e. by the law of total variance—. Instead, a proper extraction of the averaged standard deviation of these variables has been done holistically: considering the pixels of all the images as a single set and the SE and the SSIM as global variables of such set.

Ex.3 Consist in the analysis of the whole set of images in the data-set with several sets of basis-functions, i.e. at several scales (defined by $\mathbf{W} = \{W_i\}$). The first two experiments **Ex. 1** and **Ex. 2** are used to motivate a scheme for the selection of N_i^*: elbow of $M - SSIM(N)$ when following the *ranked* ordering (see discussion for details). This last experiment follows such scheme and evaluates its stability by repeating **Ex. 2** several times, one per scale value $W_i \in [3, 15]$, W_i *odd*. This experiment results in the selection of generic $\mathbf{N}^* = \{N_i^*\}$ values, quantified in Table 5.1. Results after $\mathbf{N}^*$ selection are included in Figure 5.7 in the shape of statistical surfaces. These surfaces represent the MSE and M-SSIM values for the $\mathbf{N}^* = \{N_i^*\}$ partial reconstructions of each of the images in the training set and each of the tested scales $\mathbf{W} = \{W_i\}$. Additionally, the perceptual limits of operation of the scheme are included in Figure 5.8 in the shape of the best and the worst reconstructed images.

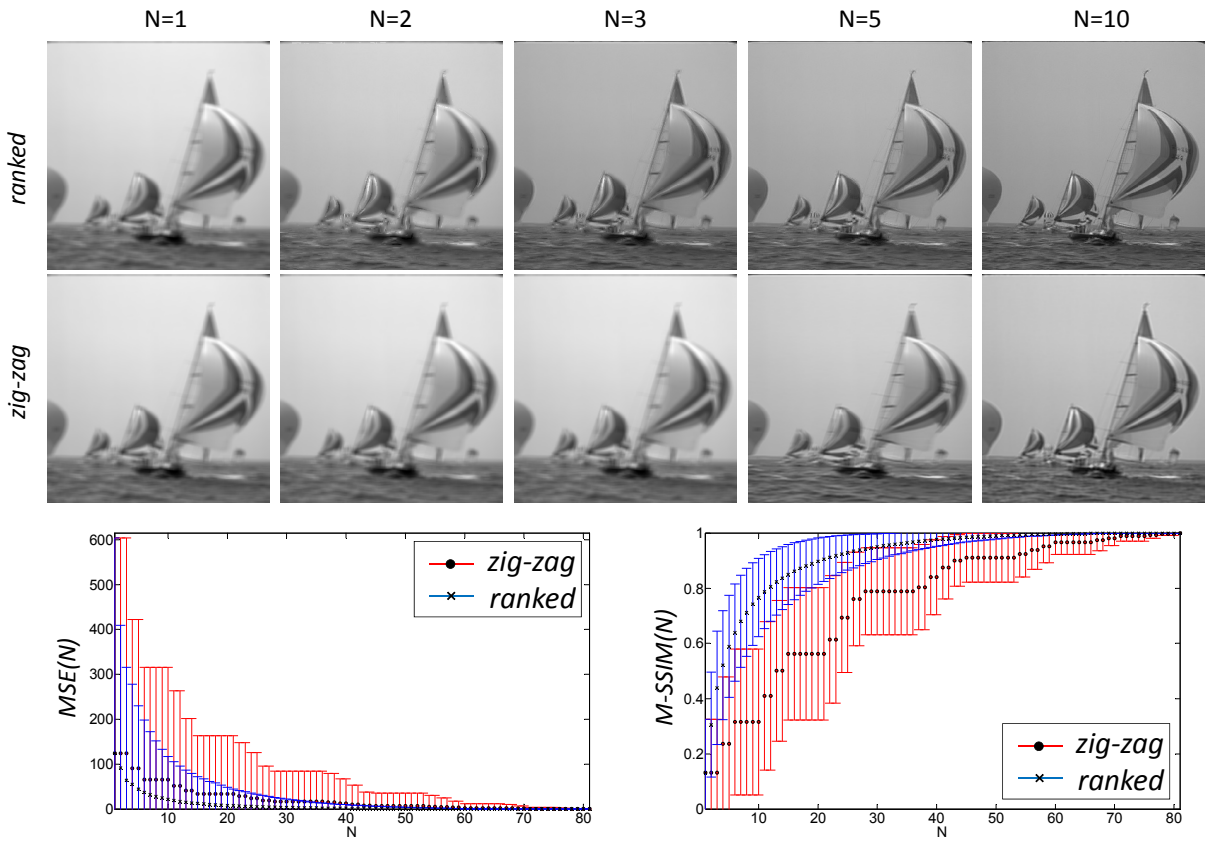


Fig. 5.4. **Ex.1.** Goodness of partial reconstruction by cutting off the DCT at the N^{th} *zig-zag* and N^{th} *ranked* ordering schemes (image example). Top row. Reconstructed images obtained by considering only the first $N = \{1, 2, 3, 5, 10\}$ AC coefficients ($W_i = 9$ in the example) according to the proposed *ranked* ordering scheme. Middle row. Reconstructions at same Ns according to the *zig-zag* ordering scheme. Bottom row. MSE reconstruction error and SE standard deviation per pixel (left) and M-SSIM index and associated SSIM standard deviation per pixel (right) for the *ranked* and the *zig-zag* ordering schemes. See text for details and discussion.

Discussion

The experiments are discussed on three basics: the prevalence of one ordering scheme over the other. The benefits of using the MSE or the M-SSIM as fidelity signals and the stability of the selected values of $\mathbf{N}^* = \{N_i^*\}$.

On the ordering schemes.

Results in Figure 5.4 show that the reconstructions obtained by following the *ranked* ordering require a lower number of coefficients to faithfully represent the image content. This can be observed in the reconstructed images in the top rows of the Figure. The DC coefficient is always the first for both ordering schemes. The DC contains the average value of each block. Hence,

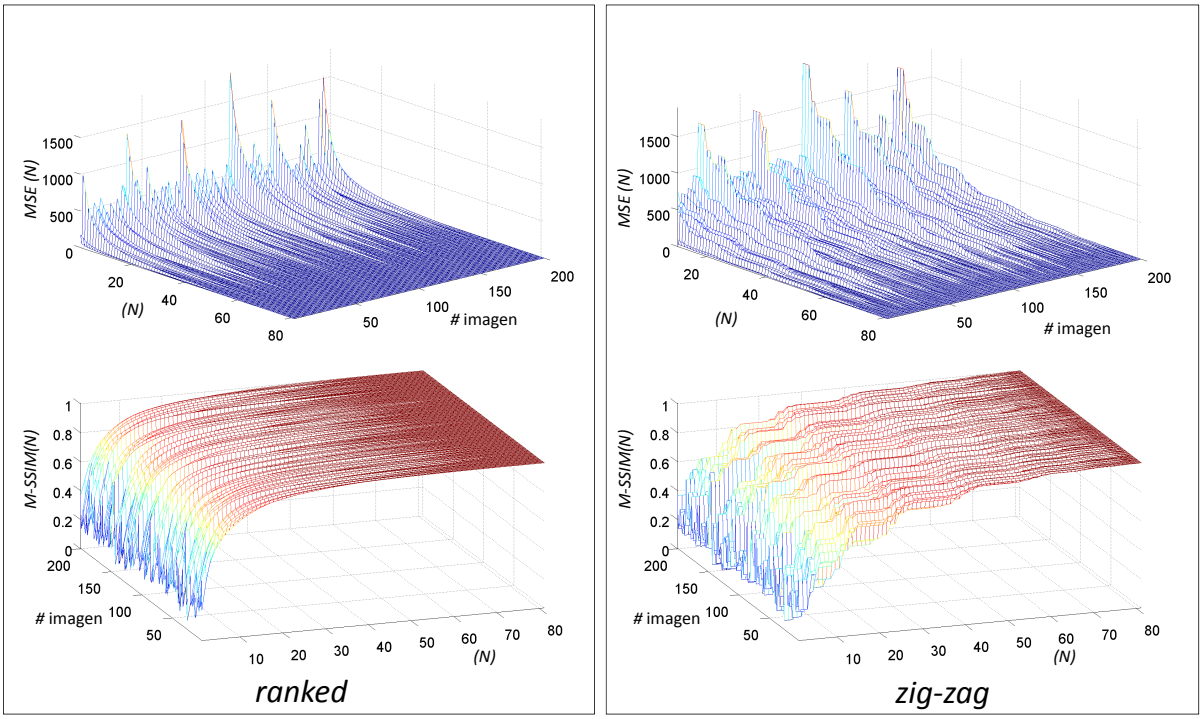


Fig. 5.5. **Ex.2 (1).** Goodness of partial reconstructions—MSE (top) and M-SSIM (bottom)—by cutting off the DCT ($W = 9$ in the example) in the N^{th} *ranked* (left) and N^{th} *zig-zag* (right) ordering schemes. Overall statistics for the whole training set. See text for details and discussion.

the image content can be perceived with just a single coefficient. However, the so-reconstructed image appears blurred and low contrasted. Reconstruction following the *ranked* ordering gets contrasted faster, i.e. requiring less coefficients, as the first in the ordering are those with more information of the image content. Fine details of the image, as the sail-boat ropes can be just—moderately—perceived in the *ranked* ordered reconstructions. The ability of first-including informative coefficients of the proposed *ranked* ordering is even clearer in the bottom row of the Figure. See, for instance the steps in the $MSE(N)$ and the $M-SSIM(N)$ obtained for the *zig-zag* scheme. These are due to the inclusion of non-informative coefficients in the reconstruction and do not appear in the *ranked* ordering. In general, the *ranked* ordering scheme achieves smaller MSE errors—the lower the better—and bigger M-SSIM values—the closer to 1 the better—for every N. Furthermore, associated standard deviation of both measures is consistently lower for every N in the proposed ordering scheme; suggesting that the majority of the image pixels are benefiting from the ordering. Exceptions are $N = 1$, to which both ordering schemes return the DC coefficient, and $N = W_i x W_i$, to which all the coefficients are used for reconstruction; hence, for these N values, both ordering schemes result in equal comparison statistics. The generality of these observations is assessed in Figure 5.5. The discussed steps

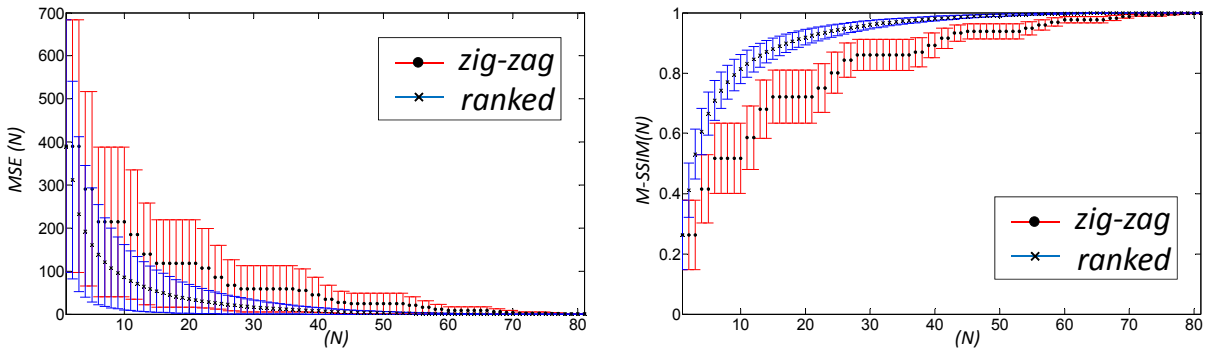


Fig. 5.6. **Ex.2 (2).** Goodness of partial reconstructions—MSE (left) and M-SSIM (right)—by cutting off the DCT ($W = 9$ in the example) in the N^{th} *ranked* (left) and N^{th} *zig-zag* (right) ordering schemes. Averaged statistics for the whole training data-set. See text for details and discussion.

in the fidelity signals can be observed in the column depicting results for the *zig-zag* scheme. Furthermore, for every analysed image, higher M-SSIM and lower MSE values are obtained for earlier Ns in the *ranked* ordering. Averaged results in Figure 5.6 are very similar to the curves in Figure 5.4 (with the exception of a decrease in the standard deviation that will be discussed later on). Hence, the same evidences derive from them: *ranked* ordering results in better fidelity signals and in smoother pure increasing (or decreasing) evolutions of these signals with N.

On the fidelity signals.

The advantages of the M-SSIM over the MSE were exhaustively discussed in Wang and Bovik [2009]. In the context of our experiments the M-SSIM is preferred for two main reasons. First, it conveys an easier and bounded interpretation of the images inter-similarity. Second, the stability of the results measured via M-SSIM is substantially higher than the stability of the results described by means of the MSE. This last issue can be observed by comparing the length of the bars representing the standard deviation of the signals in the bottom row of Figure 5.4. The same effect applies for the first column in Figure 5.5. The M-SSIM reaches stable values for every image for lower N values than MSE. The stability of the M-SSIM is specially remarkable in the second column of Figure 5.6. Comparing these deviations with those in 5.4 it is clear that the M-SSIM is benefiting from the presence of large flat image areas in the training set.

On the values and the stability of N*. In the experiments discussed up to this point, both the MSE and the M-SSIM are represented as L-shaped functions of N. A trade-off value for N, N^*, can be selected as the elbow or corner value of these curves. There are several alternatives to extract the elbow of L-shaped functions automatically. For instance, a numerical solution can rely on the second derivative of the $M - SSIM(N)$ to locate the corner or inflection point. As $M - SSIM(N)$ is a discrete variable, the second derivative can be approximated by numerical differentiation. A practical solution—probably less affected by potential inconsistencies—may

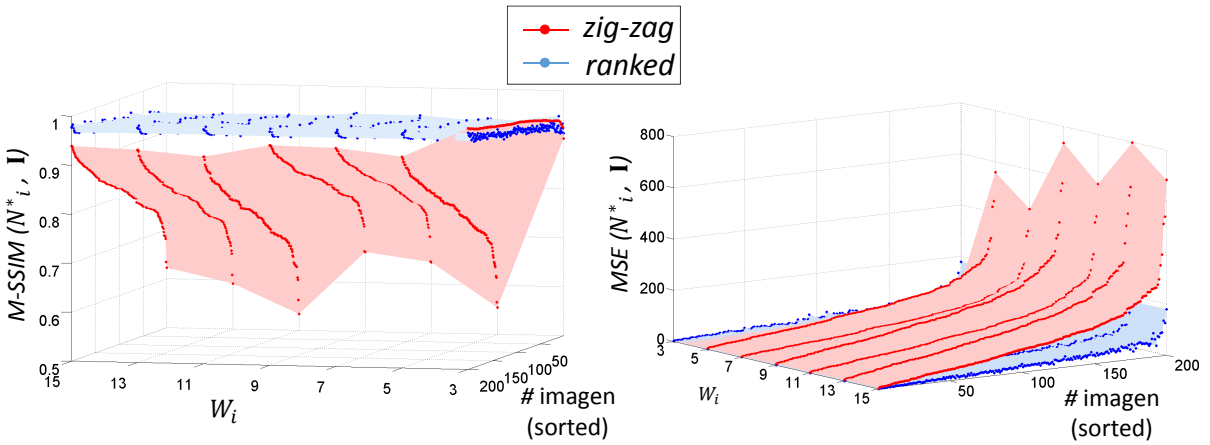


Fig. 5.7. **Ex.3 (1).** Goodness of partial reconstructions—M-SSIM (left) and MSE (right)—by cutting off the DCT in their elbow of the *ranked* and the *zig-zag* ordering schemes for different values of W_i . Overall statistics for the whole training data-set. Images has been sorted according to *zig-zag* increasing M-SSIM and MSE values to ease visualisation. See text for details and discussion and Table 5.1 for N_i^* values.

be to obtain the elbow as the N coordinate of the furthest point in $M - SSIM(N)$ to the straight line that joints the beginning of the function $[1,\ M - SSIM(1)]$ and its end $[W_i x W_i,\ M - SSIM(W_i x W_i)]$. We opt for this last approximation to obtain the values of $\mathbf{N}^*$ consigned in Table 5.1. The M-SSIM values for these cut-off numbers reinforce the idea that the *ranked* provides better reconstructions than the *zig-zag* ordering by using a lower number of basis-functions responses.

Figure 5.7 is devoted to evaluate the stability of the selection $\mathbf{N}^*$ for different scales and images. The *ranked* ordering result in stable measures of the M-SSIM (see the *quasi*-flat statistical surface created in the left column of the Figure). Note that, this implies that with independence of the image analysed, the proposed method for the selection of the relevant basis-function following the proposed *ranked* ordering achieves M-SSIM values which ranges from 0.765 (the worst) to 0.982 (the best). Reconstructions of these extreme images are included in Figure 5.8.

The designed scheme allows to use a predetermined number of basis-functions (the nature defined by the *ranked* ordering) given a DCT scale with independence of the image. However, as suggested in section 5.3 the scheme can be instead apply on each particular image to improve the fidelity of the selection. We opt for the predetermined values and, from here in advance use the $\mathbf{N}^*$ values in the eight row of Table 5.1 to configure the local-variability description method.

Fidelity	Ordering	W_i:	3	5	7	9	11	13	15
MSE	*zig-zag*	N_i^*	6	6	15	15	28	28	45
		error (N_i^*)	**14.07**	88.14	69.23	117.9	94.94	129.4	108.5
	ranked	N_i^*	4	7	10	15	19	25	31
		error (N_i^*)	27.18	**38.32**	**51.64**	**53.66**	**62.56**	**65.17**	**68.98**
M-SSIM	*zig-zag*	N_i^*	6	6	15	26	28	45	64
		error (N_i^*)	**0.966**	0.793	0.838	0.843	0.772	0.803	0.816
	ranked	N_i^*	4	6	10	16	22	30	39
		error (N_i^*)	0.945	**0.903**	**0.891**	**0.890**	**0.876**	**0.870**	**0.864**

Table 5.1: . Optimal number of DCT filters for different block-size values (scales). Value of N at elbow: N_i^* and associated value of the fidelity measures for several scales values. Best figures per scale and fidelity signals are highlighted in **bold**.

5.6 Building a contour map.

We propose to define the likelihood of a pixel $\mathbf{p}$ being part of a contour by comparing it with the $W_i x W_i-$neighbourhood around it: $\mathcal{N}_{\mathbf{p},W_i}$. The comparison is performed in terms of the pixel's N_i^*-truncated local-variability descriptions (the coefficients and basis-functions in equations 5.8 and 5.9) obtained at a given scale W_i . The comparison is measured by means of the pixel-wise distance in equation 5.15. The distance to each of the neighbours is weighted by a rotationally symmetric Gaussian kernel of standard deviation equal to the scale, W_i centred at $\mathbf{p}$. Per-scale likelihoods are aggregated following the scheme in equation 5.17 to finally derive a contour likelihood for each pixel, $CM(\mathbf{p})$—the higher the more likely $\mathbf{p}$ being part of a contour—:

$$CM(\mathbf{p}) = \sum_{i}^{|\mathbf{W}|} \sum_{\mathbf{p}' \in \mathcal{N}_{\mathbf{p},W_i}} g(\mathbf{p}'; W_i) \cdot d_{L-V}(\mathbf{p}, \mathbf{p}') \tag{5.18}$$

Discussion of preliminary results

In Figures 5.9, 5.10 and 5.11 we include example contour maps extracted on images in the test set of the BSD500 data-set. In these maps, it can be observed that the proposed method is able to detect contours when these occur between two flat areas or between a flat and a textured area. However, the proposed method present some problems when dealing with transitions between two different textured areas. This effect suggest that further research should be made in the use of the responses intensities when comparing pixels (see equation 5.16). Nevertheless, preliminary results suggest that the method is able to handle textured areas (see examples in the grass and the straw of the example images). Finally, probably the biggest problem in the

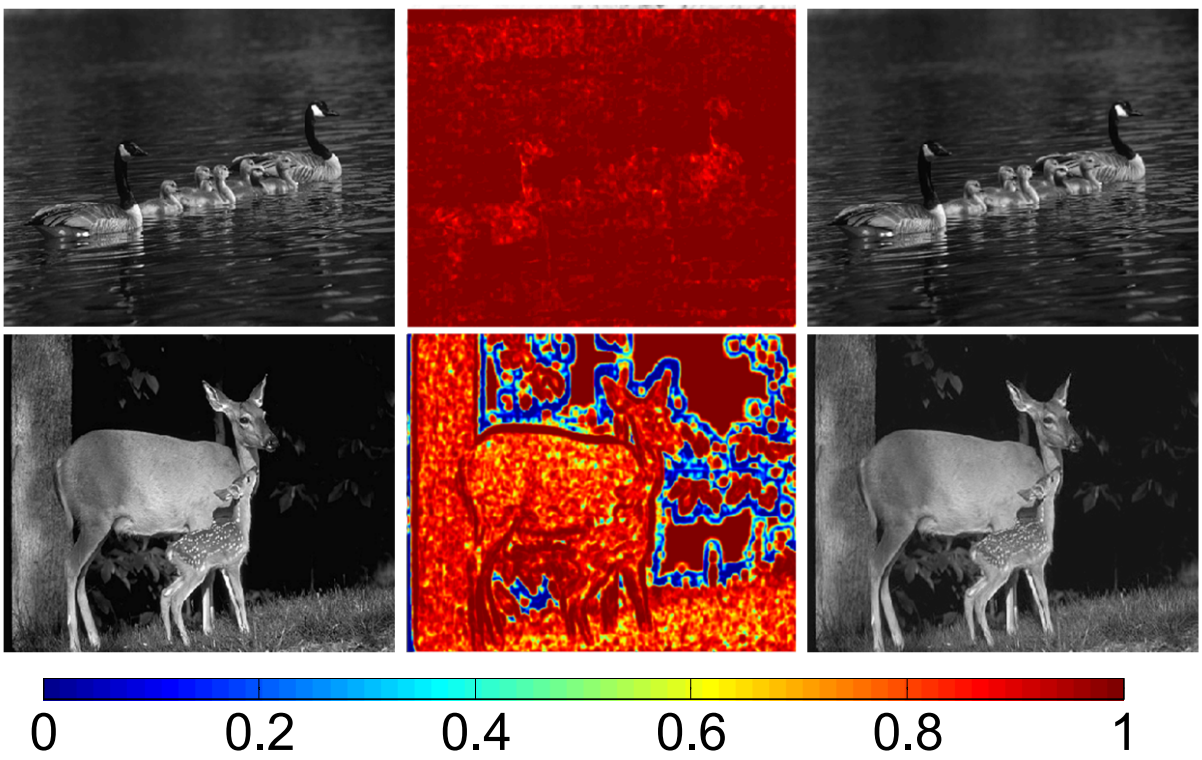


Fig. 5.8. **Ex.3 (2).** Best (top row) and worst (bottom row) reconstructed images in the training set. Left column. Original luminance images. Middle column. SSIM values per pixel. Right Column. Reconstructed images. Best reconstructed image is obtained for $W_i = 3$ ($N_i^* = 4$) and results in a M-SSIM value of 0.982. Worst reconstructed image is obtained for $W_i = 15$ ($N_i^* = 39$) and results in a M-SSIM value of 0.765. Note that errors are condensed in the dark areas of the image. Nevertheless, just small perceptual differences can be perceived between the reconstructed and the original image.

maps is the incomplete nature of the object contours (see examples in the arch and the plane in Figure 5.9, in the lizard in Figure 5.10 and in the house roof in Figure 5.11). This indicate that using solely the contour map might be insufficient for the detection of image transitions (the same reflection was implicitly raised in Martin et al. [2001]; Arbelaez et al. [2011]), our future work will be devoted to develop the combination of the information in the contour map with additional features.

5.7 Chapter conclusions.

In this chapter we have proposed a Texton-based discriminative method to describe local-variability in natural images. The method relies on the ability of the Discrete Cosine Transform to provide an easily parametrisable filter-bank on which multi-scale analysis can be also easily

configured. Responses of an image to the filters in the filter-bank are analysed pixel-wise and only a subset of these responses are used considered useful cues. This subset is decided in terms of the representativeness of the responses. An experimental analysis of such relevance in a set of varied images has led to the interesting reflection that the number of relevant responses can be selected the same for any pixel with independence of the image if small errors are tolerated. So-built descriptions were used to detect contours by comparing the characterisation of adjacent pixels. To this aim, we designed a metric based on the subjective comparison of the spatial patterns representing the filter in a DCT filter-bank. Finally, we propose to use both the selection scheme and the metric to derive a contour map whereby encode the likelihood of a pixel being part of a contour. Preliminary results suggest that the use of this contour map in combination with additional features may provide a suitable scheme to detect object contours in an image.

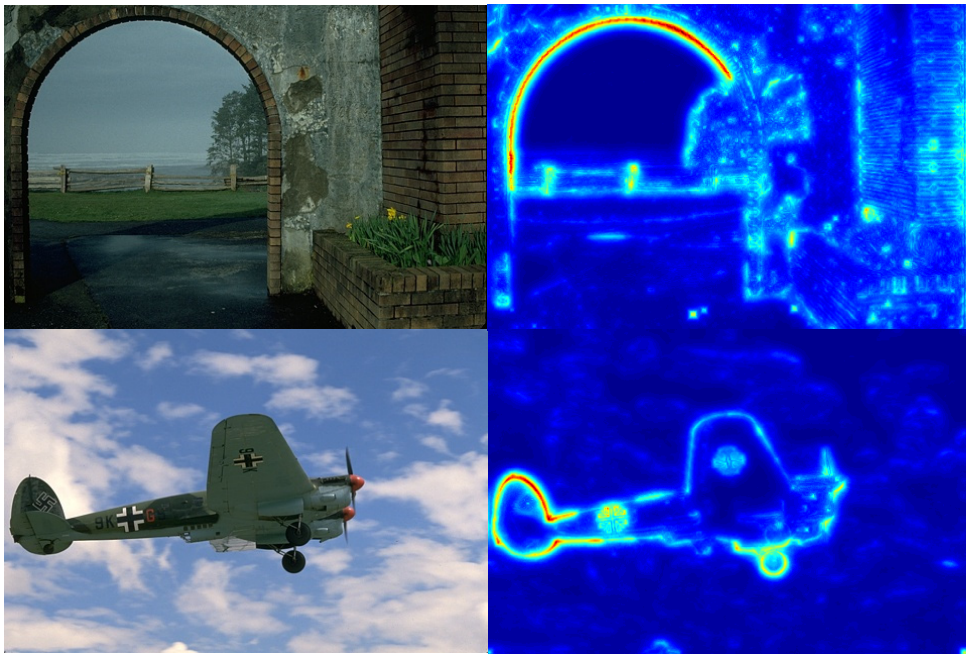


Fig. 5.9. Contour map examples. Original RGB images (left), extracted contour maps (right). The redder (the bluer) the higher (the lower) the likelihood of a pixel being part of a contour.

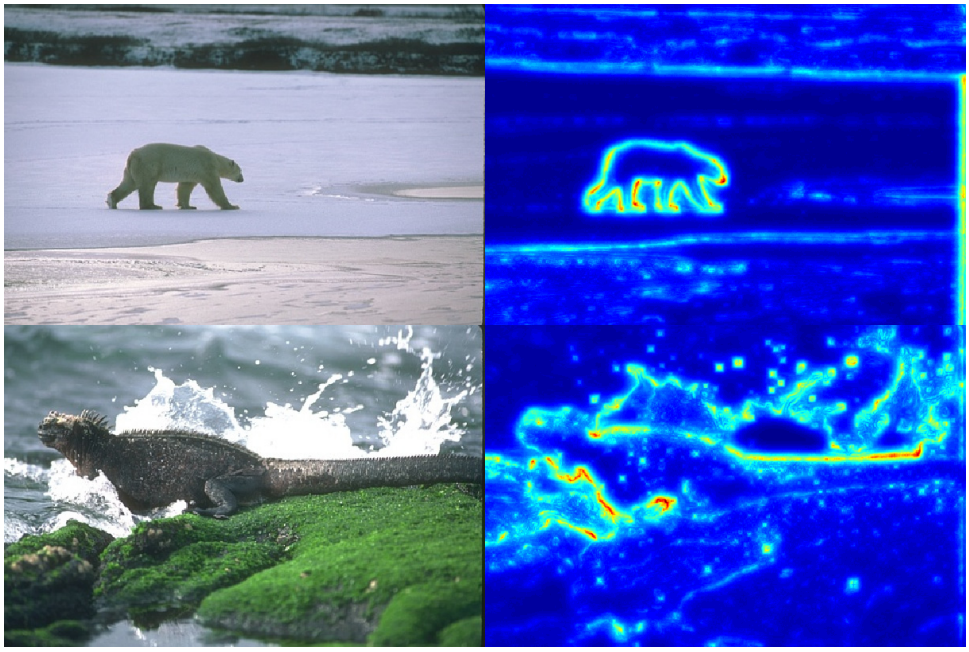


Fig. 5.10. Contour map examples. Original RGB images (left), extracted contour maps (right). The redder (the bluer) the higher (the lower) the likelihood of a pixel being part of a contour.

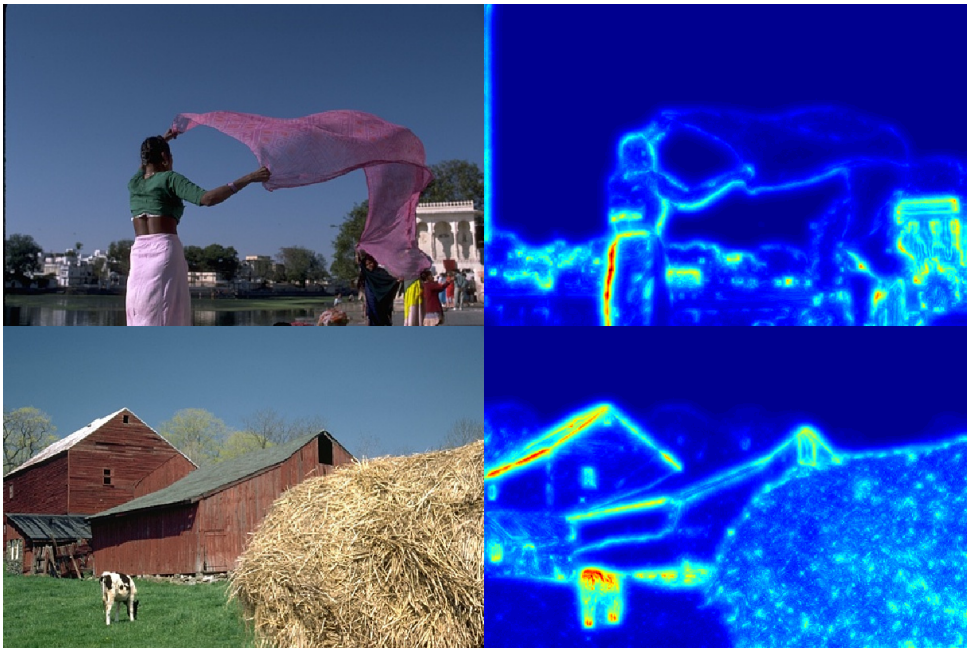


Fig. 5.11. Contour map examples. Original RGB images (left), extracted contour maps (right). The redder (the bluer) the higher (the lower) the likelihood of a pixel being part of a contour.

Part III

Part III. Regions for background subtraction

Contents

This part studies the use of region-driven schemes as complementary methods to pixel-based background subtraction.

The part starts in chapter 6 with a review of background subtraction approaches in terms of their stages of processing and emphasising the associated challenges that background subtraction entails. We propose to organise existing methods on a per challenge basis. Indicating the solutions and key proposals to face them. The chapter ends with a brief discussion on existing data-sets and metrics for evaluation. In chapter 7 we sketch a couple of region-driven methods to refine results of a pixel-based noisy algorithm. The first proposal aims to overcome problems caused by local-illumination changes by extrapolating pixel-based results to illumination-stable regions. The second proposal defines a whole background model based on the use of regions and present a covariance-based comparison framework to model and store the dynamism of multi-modal backgrounds with a generic number of description features. The contents in this part are completed with Appendixes A and B.

"For me, a landscape does not exist in its own right, since its appearance changes at any moment."

Claude Monet (a founder of French impressionist painting)

Chapter 6

Challenges, key-tasks and recent trends in background subtraction

In scenarios recorded by a static camera, the problem known as Background Subtraction (BS), i.e. the problem of automatically segmenting each frame in relevant (foreground) and irrelevant (background) objects—as a two-class problem, it is also sometimes known as background-object segregation—has been recursively studied. Although BS is not the only technique available for this task—alternatives include motion-compensation (Neri et al. [1998]; Jain et al. [2013]) and image-scanning approaches (Felzenszwalb et al. [2010b])—, it has been, by far, the most used and referenced.

The problem can be stated as it follows:

Let $\mathbf{I}$ be a particular frame of a video recorded by a fixed camera. BS approaches search for a division of $\mathbf{I}$ into a set of foreground pixels $\Omega_F(\mathbf{I})$ and a set of background pixels $\Omega_B(\mathbf{I})$:

$$\mathbf{I} = \Omega_F(\mathbf{I}) \cup \Omega_B(\mathbf{I}), \text{ such that } \Omega_F(\mathbf{I}) \cap \Omega_B(\mathbf{I}) = \emptyset \tag{6.1}$$

, i.e. for a region segmentation with only two labels and without requiring the regions to be connected components (see Chapter 2).

There is a significant quantity of scientific studies that use BS as a primary tool to feed higher-level tasks, including: object tracking, object/people recognition or scene understanding. This multi-task nature leads to two major implications: i) BS has been widely used in many computer-vision applications such as video surveillance, traffic monitoring and human computer interfaces and ii) BS has been exhaustively studied—with up to 160.000 entries in Google Scholar—. The principle of BS algorithms is to build a model of the empty scene (commonly named as background) and then detect—and segregate—objects of interest as elements (usually called foreground) that do not fit into the background model. According to Bouwmans [2014], a BS algorithm can be described by its solutions to the following key-tasks:

Background initialisation: defines the strategies to initialize the background model, ideally with a video frame free of foreground objects thus determining an appropriate point of departure for the background modelling stage.

Background modelling: describes the nature of the background model and its associated statistics used to store the empty scene—this task is also known as *background representation*—.

Background maintenance: this task is devoted to adapt the background model to the changes occurred in the scene over time.

Foreground detection: in this task the *difference* between incoming video frames and the background model is evaluated according to a set of features.

This chapter briefly summarises existing approaches to confront the task of BS. We propose to organise them on a per-stage basis remarking the challenges they intend to resolve. To this aim, we first review the common challenges that should be faced when designing a BS approach. Then, we extend the definition of each of the key-tasks defined and describe both top-performing approaches and classical methods according to this organisation. Following, we describe the data-sets and the metrics used for benchmarking of BS approaches. The chapter ends with a discussion about the lack of region-based approaches and with brief conclusions on the topic.

6.1 Challenges in BS

According to the challenges in BS, those identified by Toyama (Toyama et al. [1999]) are still the reference. Furthermore, in Bouwmans [2014] three new camera-related challenges are included. We propose to organise them in three categories, according to the challenge's source: camera, background and foreground—challenges caused by several sources, as camouflage, are here assigned to foreground—. These can be listed, slightly modifying the nomenclature in Bouwmans [2014], as:

Camera-related challenges

- Image noise: includes the acquisition-noise in the recording process, the interpolation-noise of resized frames and the block-noise of decompressed videos.
- Camera jitter: when static cameras are placed in non-stable supports—as highway's cameras placed on bridges or poles—wind can make the camera vibrate, resulting in nominal motion and—if uncompensated—, in false foreground detections.

- Camera automatic adjustments: the automatic processes included in some cameras to adapt to scene changes—including refocus, automatic control gain, white balance and brightness control—may completely change the background appearance respect to that modelled.

Background-related challenges

- Illumination changes: these can be divided into global and local. Global can be further subdivided into gradual—daylight in outdoors scenes—and abrupt—switch on and off of lights in indoors scenarios. Local changes include self-shadows and highlights.
- Removed background objects: inanimate background objects can be taken—e.g. stolen—by animated foreground objects—e.g. a person—, leaving a wake—also known as a *ghost*—in the original position.
- Inserted background objects: the opposite of removed background objects; inanimate objects may be placed in the background. Both situations are especially common in surveillance scenarios.
- Dynamic backgrounds: especially problematic in outdoor scenarios where some parts of the background may be moving. This motion might result in large difference—multimodality—respect to a simple background model. Common examples of dynamic backgrounds include moving water and waving trees.

Foreground-related challenges

- Bootstrapping: in crowded scenes part of the background can be occluded for a long time, then hindering the availability of enough samples to model its evolution or even its appearance.
- Shadows: whereas background shadows—self-shadows—can be considered an illumination issue, foreground or moving shadows—commonly named *cast-shadows*—represent a problem as they move as foreground while being represented by similar but lower-intense modes than those in the background model.
- Beginning moving object: it is sometimes considered the equivalent of a removed background object for foreground objects. The main difference relies on the object nature—*animated* objects—: usually people, moving cars or animals.
- Sleeping foreground object: The parallelism continues with this human-driven version of inserted background objects. Even though the decision of incorporating these objects to the background model—a decision that potentially leads to beginning moving object

situations—or not is task-dependent, people is usually expected to move again. If the foreground object is there since the initialization—and no management of this situation is performed—the challenge is also known as a *hot-start*.

- Camouflage: it is probably—together with bootstrapping—the least studied challenge. Background and foreground objects may share similar—or even equal—appearances, then leading to an inaccurate discrimination process. Obviously this challenge is feature-dependent.

- Foreground aperture: this challenge only applies for homogeneous foreground objects that were incorporated to the background model. Partial movements of these objects are only detected on the boundaries whereas the interior remains equal to the stored appearance in the background model. In our opinion, it is a special sequence of three challenges: sleeping foreground object, beginning moving object and camouflage. However, foreground aperture may be also explained by the sequence: removed background object and camouflage. For both cases the consequence is foreground miss-detection.

Despite the enormous amount of efforts and studies devoted to solve them, the research community agrees (Toyama et al. [1999]; Elhabian et al. [2008]; Cristani et al. [2010]; Bouwmans [2014]) that it does not yet exist a system able to solve all of these challenges at the same time. This is mainly due to a tug-of-war between generalist background-modelling and accurate foreground detection; i.e. enhancing approach's flexibility to learn the different background appearances usually harms its ability to adequately discriminate the foreground.

Furthermore, they cannot be solved in the same stage of processing; those related to illumination changes need to be addressed in the modelling and updating stages, and those associated with the foreground density, e.g. bootstrapping, usually require also specific solutions in the initialization stage. Camouflage is usually ignored or managed in model-blind post-processing stages relying on colour and luminance features. In our opinion, it should be better managed in the foreground detection stage by exploring new features and metrics (as in St-Charles et al. [2015]).

6.2 Key-tasks and relevant trends in BS

Figure 6.1 illustrates a generic scheme to describe BS approaches. Let us review existing solutions on a per-task basis, identifying the BS challenges faced by each task.

Background initialisation

The solutions undertaken in this task define the strategies to initialise the background model. It is an important stage of the BS algorithms that has been weakly investigated in comparison with

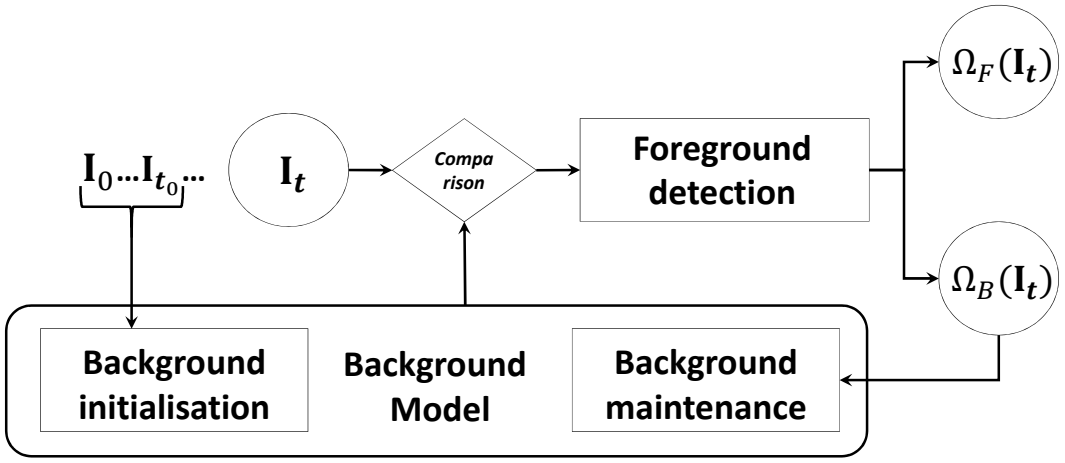


Fig. 6.1. Generic flowchart to describe background subtraction approaches. The t sub-index in $\mathbf{I}_t$ represents the temporal dimension.$\mathbf{I}_0$ is the first frame of the video and $\mathbf{I}_{t_0}$ the last one used for the initialisation, i.e. for on-line initialisation approaches $t_0 = 0$.

the others (Bouwmans [2014]). It consists of initialising the background model by extracting a background image free of foreground objects. The complexity of this stage lies in the presence of several challenges in the beginning of the sequence—bootstrapping, camouflage, stationary objects and illumination changes and shadows—. There are three main strategies to initialise the background: on-line—the background is built with the frame's temporal evolution (Colque and Camara-Chavez [2011]; Crivelli et al. [2011]; Zhang et al. [2012]; Hsiao and Leou [2013])—, batch—completely analyses a training sequence to generate the background (Wang and Suter [2006]; Colombari and Fusiello [2010]; Colque and Camara-Chavez [2011])—and hybrid—in spite of relying on batch-like techniques, is able to deliver a background at each temporal instant (Baltieri et al. [2010]; Reddy et al. [2011])—.

Batch and hybrid approaches, whereas able to obtain better backgrounds than on-line initialisations, are either infeasible to operate without knowing the whole sequence (batch) or are too-dependent of previous data (hybrid). The former requirement inhibits its use in on-line demanding applications (e.g. surveillance scenarios) whereas the latter dependence implies a problem when substantial and global changes occur in the video. For these reasons, on-line initialisation is preferred in the majority of the existing approaches

Background modelling

The background modelling stage has classically been the main criteria to organise BS approaches. In fact, for years, BS approaches were divided into parametric—evolutions of the well-known Mixture-of-Gaussians MoG (Stauffer and Grimson [1999]; Evangelio and Sikora [2011]; Evangelio

et al. [2014]; Wang et al. [2014a])—and non-parametric—a successful alternative (Barnich and Van Droogenbroeck [2011]; Hofmann et al. [2012]; St-Charles et al. [2015])—which includes evolutions of the top-referenced Kernel-Density-Estimation KDE (Elgammal et al. [2000, 2002]) and is currently the preferred line of research. Note that non-parametric approaches range from those aiming to estimate the density of the data distribution—e.g. the KDE Elgammal et al. [2002] itself or Zivkovic and van der Heijden [2006]—to those storing a determinate number of samples which inherently describe the density without estimating it—e.g. solutions in Barnich and Van Droogenbroeck [2011]; St-Charles et al. [2015]—.

Cluster and codebook models (Butler et al. [2005]; Kim et al. [2005]) and PCA-based subspace-learning models (Bouwmans [2009]) are the rare alternative to these two major schemes. Recently, we can also identify new trends in background modelling, including: modelling based on self-organised neural networks (Maddalena and Petrosino [2012]), uncertainty-based fuzzy models (Kim and Kim [2012]) and evolutions of sub-space methods (Bouwmans and Zahzah [2014]). However, their degree of success is still far from recent non-parametric methods.

Aside from this parametric vs non-parametric discrimination we can also distinguish two main schemes: mono-layer and multi-layer. On one hand, mono-layer approaches accumulate background statistics on a single layer. These statistics can be modelled by simple or complex schemes but every incoming sample to be modelled affects the whole model. On the other hand, multi-layer approaches use several layers to model the background—and occasionally foreground—statistics. Layers are usually devoted to store the different appearances of the background—e.g. due to captured noise or dynamic backgrounds—. The main advantages of using multi-layer schemes are: (i) modifications of the (sub)models in each layer do not affect the rest of the layers, and (ii) the likelihood of a sample belonging to a layer, and of a layer belonging to the background are in general independent. This last advantage implies that a layer can be modelled without making prior decisions about the nature of the samples that model it—either foreground or background up to this point—. In a posterior stage of analysis the layer can be further classified according to their temporal evolutions. However, multi-layer schemes usually require more complex updating processes and foreground detection schemes than their mono-layer counterparts.

A different scheme is used by each of the two top-performing approaches to date. The BS approach described in Wang et al. [2014a] relies on a parametric single-layer model which operates by extending MoG incorporating motion information and adapting the number of required Gaussian with the video content. The algorithm presented in St-Charles et al. [2015] instead opts for a non-parametric modelling storing up to 50 different appearances for each background sample.

Basically, all these methods aim to provide flexible and adaptable models able to handle dynamic backgrounds.

Background maintenance

Usually closely linked with the background modelling stage, maintenance mechanisms are fully included in the model definition. However, we can distinguish two main strategies for model maintenance: blind and selective. On one hand, blind maintenance mechanisms equally consider every incoming sample—both background and foreground samples—for updating the model. On the other hand, selective maintenance approaches use different strategies for background samples—usually some sort of running average scheme—than for foreground samples—most of the times discarded—.

Evolutions of these strategies include multi-class selective updating (Colmenarejo et al. [2011]) and confidence-driven updating (Porikli and Tuzel [2005]). On one hand, multi-class updating enhances the segregation process (foreground-background) by introducing gradual classifications —e.g. foreground-shadows-background at different stages of modelling—and designing *ad hoc* maintenance strategies for each class. On the other hand, confidence-driven updating combines the likelihood between new background and model samples with the temporal evolution of the background samples to adapt the learning rate. Appendix B includes and example of a BS algorithm that combines both strategies.

However, alternative updating schemes have been also recently presented. For instance, a successful scheme relying on a dual modelling of the background evolution is proposed in (Evangelio et al. [2014]) . The strategy is based on maintaining two background models with different updating ratios. These are named short and long term models. The long term model can be used to inhibit or allow updating of the short term modelled statistics.

An additional scheme is the storing of statistics of the foreground detection results on each pixel. This line of research, known as *self-adaptation* or *self-tuning* has been intensively explored recently (Hofmann et al. [2012]; Wang and Dudek [2014]; St-Charles et al. [2015]). In general terms, these approaches use statistics about the quality of the model when identifying foreground samples and use these to update the models accordingly. The statistics range from the increasing/decreasing of the matching distance to the model, to the monitoring of blinking pixels, i.e. those in which foreground is alternatively detected.

Finally, in non-parametric background models—as the storing capacity is constrained in digital systems—two schemes have been defined for sample replacement: random-substitution (e.g. Barnich and Van Droogenbroeck [2011]; St-Charles et al. [2015]) and based on the description capacity of the stored samples (Zivkovic and van der Heijden [2006]), the lower the sample description capability, the higher its likelihood of being replaced.

Robust maintenance mechanisms used in flexible models aim to overcome camera—noisy image, camera automatic adjustments—and background—illumination changes—related challenges. Furthermore, is in this stage where the maintenance mechanism define whether inserted objects are incorporated to the model—inserted background objects, sleeping foreground

object—and whether ghosts—removed background objects, beginning moving object—are updated.

Foreground detection

Foreground is detected as unobserved—or not explained by the model—samples. However, whereas this is sometimes understood as a simple classification task (Bouwmans [2014]), in our opinion it is a key task for obtaining accurate results. Detection is mainly driven by the features used for characterising the samples and by the distance used to measure the separation among these characterisations.

On one hand, several features Li et al. [2004] have been proposed in the literature, with colour and luminance—also known as *spectral* features— being the favourite option (Bouwmans [2014]). Spectral features operate well in most scenarios but suffer from camouflage, shadows, foreground aperture and illumination changes. Alternatively, texture features can be used to remove ghosts and are assumed to operate better where colour fails. Among the texture features, those which extraction requires small amount of processing are preferred: Local Binary Patterns (LBP) in Heikkilä and Pietikäinen [2006], or extended versions of LBP as the Local Binary Similarity Pattern (LBSP) in St-Charles et al. [2015]. Finally, disparity and depth—i.e. obtained by stereo reconstruction (Ivanov et al. [2000]) or by *Time-of-Flight* cameras (Molina et al. [2013])—are considered the best features for handling camouflage and illumination-related challenges, but require the use of at least two cameras recording the scene. On the other hand, among the comparison strategies there are few alternatives rather than different kind of norms—non-parametric—and responses to the estimated models—parametric—. However, schemes including covariance-driven comparison have been also proposed (Zhang et al. [2008b]). Spectral features can be also used to compute second-order features as motion—which inherits the advantages and disadvantages of the feature(s) used to obtain it—(Wang et al. [2014a]). An alternative is to combine several features through different schemes: Zhang and Xu [2006]; Bhaskar et al. [2010].

6.3 Evaluation of background subtraction approaches.

The proper evaluation of BS approaches was initially very complex due to the existence of several small data-sets with few scenarios and using different ground-truth annotations. However, in 2012, the Change Detection data-set (Goyette et al. [2012]) was proposed to handle this issue. The data-set encompasses 6 scene categories: *baseline*, *dynamic background*, *camera jitter*, *Intermittent object motion*, *shadow* and *thermal*. Along these categories several challenges are

inspected and properly annotated. The data-set was extended in 2014 (Wang et al. [2014b]) including 5 new categories: *bad weather*, *low frame-rate*, *night videos* and *PTZ.*

Concerning the evaluation statistics, BS is considered as a classification approach and then classical expressions combining the true positive (TP) false positive (FP), true negative (TN) and false negative (FN) indicators are used to quantitatively measure the BS performance. Aside from recall (R), precision (P) and $Fscore$, which were already defined in chapter 3, four other metrics are typically used: specificity $Sp = TN/(TN + FP)$, false positive rate $FPR = FP/(FP + TN)$, false negative rate $FNR = FN/(TP + FN)$, and percentage of wrong classifications $PWC = (FN + FP)/(TP + FP + FP + TN)$.

6.4 Discussion.

Whereas the existence of the Change Detection data-set provides a proper corpus for evaluation it may occlude the low level accuracy of BS methods, as the performance statistics ignore the error distributions. Let us explain this by an example. An accurate method producing a tightened-to-ground-truth foreground mask would be severely penalised if, due to noise, some small areas in the background are mainly classified as foreground in several frames. Therefore, a post-processing refinement of the foreground binary mask—mainly carried out by morphological operations—is usually performed to improve the algorithm performance statistics. However, this post-processing usually affects the shape of the foreground masks, deforming them and turning them less accurate and thus of a lower utility for hypothetical posterior applications in the analysis path. The problem is then in deciding whether to enhance results statistics or to enhance results utility. Similar problems related with current evaluation methods are further inspected and discussed in Margolin et al. [2014].

Furthermore, the existence of this keystone data-set also entails a significant problem: approaches adaptation. The solutions proposed in recent approaches appear to be highly influenced by the data-set nature. For instance, the current leading approach (St-Charles et al. [2015]) presents severe problems in the handling of dynamic backgrounds—which was explicitly the challenge that has motivated the higher number of solutions in the past—. However, as the number of sequences containing dynamic backgrounds represents a moderate percentage of the total number of sequences in the data-set (6 out of 53) the system just achieves a 5th position in this category while still leads overall results.

A particularly relevant observation for our study is that none of the referenced methods use regions to operate. This is mainly due to the high increase of the computational complexity and processing cost that regions entail. BS approaches—being a preliminary analysis stage of several higher-level applications and a key processing stage in surveillance scenarios—cannot tolerate this cost increment. In practice, this inhibits the proper evaluation of region-based approaches

as the cost of processing a huge amount of frames is really high. Nonetheless, we still believe that future improvements in region segmentation as well as future development of processing hardware will make suitable the use of regions for background subtraction.

6.5 Chapter conclusions.

The huge amount of existing BS approaches hinders their proper organisation. We have adopted the organisation proposed in a recent survey Bouwmans [2014] enhancing it by including recent solutions. Furthermore, we have described the challenges a BS solution must face and we have associated them with each stage of analysis. We have finally presented the agreed benchmarking solution and associated metrics for evaluation and we have discussed the problems that comparative evaluation via this corpus entails. Finally, we have briefly discussed the reason why region-based solutions are scarce in the state-of-the-art in BS.

Chapter 7

Contributions to region-driven background subtraction

In chapter 6 we have reviewed the more relevant and successful approaches for the task of BS. None of them relies on regions in any stage of their analysis. We have discussed that this absence of relevant methods is mainly due to the substantial computational-cost increase that region segmentation entails.

Despite this problem, we still think that regions can be useful in some stages of analysis mainly due to their ability in aggregating similar pixels and their ability to expand cues from pixels assigned a reliable result to hesitant-results pixels. In this chapter we describe two approaches to exemplify the use of regions for BS and illustrate how pixel-based approximations may take advantage of their potential benefits for the detection and refinement of foreground pixels. Whereas the operation of these proposals requires further evaluation, their preliminary results are promising.

The first proposal is a simple post-processing approach to easily detect shadowed as well as especially lit areas—known as *highlighted* areas—without the use of complex thresholds nor the requirement of conversion to an alternative colour space. The second proposal enhances the first one by providing a flexible and robust framework for general region characterization and matching in a BS scope. The first proposal is described in section 7.1 and the second proposal in section 7.2. Overall conclusions are included in section 7.3.

7.1 Case of example 1: illumination-blind regions for BS

Main idea and motivation.

As discussed in chapter 6, most of the existing approaches model the background by means of a mixture of statistical models or by accumulating background samples, in order to perform, via

BS, a discrimination between foreground and background pixels. However, situations such as the presence of sudden changes in illumination or the presence of moving shadows, are problematic. The problem is that these illumination effects result in frame areas which are conceptually part of the background scene but present changes respect to the model that are comparable to those produced by the foreground. Existing solutions to these problems usually rely on post-processing algorithms explicitly designed to handle these illumination issues. We instead claim that regions are not only a natural way to ascend from pixel analysis to object segregation (see chapter 2), but also a key intermediate step to control pixels aggregation while considering illumination issues. The objective of the approach described in this section is to refine a pixel-based segmentation-mask produced by a simple BS technique, via applying simple and efficient illumination constraints at region-level.

Problem statement.

Let us here review the aim which served us to introduce BS in chapter 6. Being $\mathbf{I}$ a particular frame of a video recorded by a fixed camera, BS approaches search for a division of $\mathbf{I}$ into a set of foreground pixels $\Omega_F(\mathbf{I})$ and a set of background pixels $\Omega_B(\mathbf{I})$:

$$\mathbf{I} = \Omega_F(\mathbf{I}) \cup \Omega_B(\mathbf{I}), \text{ such that } \Omega_F(\mathbf{I}) \cap \Omega_B(\mathbf{I}) = \emptyset \tag{7.1}$$

This section especially focuses on the problematic related to shadows and to strong increases of illumination (highlights). Let us focus just on shadows at this point. Shadows (both cast shadows from foreground moving objects, and shadows that vary due to small displacements of background objects) may produce image intensity variations comparable to those caused by moving objects, hence being frequently classified as foreground. As shadows are inherent to the presence of objects, much effort has been devoted to either detecting them or minimising their effect by further dividing the subset of foreground pixels into shadow pixels and objects:

$$\Omega_F(\mathbf{I}) = \Omega_S(\mathbf{I}) \cup \Omega_{F*}(\mathbf{I}), \quad \Omega_{F*}(\mathbf{I}) \cap \Omega_S(\mathbf{I}) = \emptyset \tag{7.2}$$

, where $\Omega_S(\mathbf{I})$ stands for the set of shadow pixels.

In order to discriminate these pixels from foreground pixels, some works empirically tune sets of thresholds based on rules in the HSV colour-space (Prati et al. [2003]). We consider that these mainly empirical approaches, while practical and successful for some types of video, ill-define the tasks of segmentation and tracking.

On the contrary, the approach proposed here relies on a practical approximation to the illumination model used in Nayar and Bolle [1996]; Nadimi and Bhanu [2004]:

$$L = \int \varsigma(\Lambda) e(\Lambda) r(\theta, \upsilon, \mathrm{n}, \Lambda) d\Lambda \tag{7.3}$$

, where L is the image brightness value captured by a camera sensor with spectral response $\varsigma(\Lambda)$, assuming an illumination source with a spectral distribution $e(\Lambda)$ that emits over an object surface with an angle θ respect to its normal vector n. The distribution of the reflected light can be described by the reflectance function $r(\theta, \upsilon, \mathrm{n}, \Lambda)$ with, υ, the camera viewing angle.

Our aim is to obtain a region-based frame segregation in which shadow pixels are assigned to background pixels:

$$\mathbf{I} = \Omega_{F*}(\mathbf{I}) \cup \Omega_{B*}(\mathbf{I}), \quad \Omega_{B*}(\mathbf{I}) = \Omega_{B}(\mathbf{I}) \cup \Omega_{S}(\mathbf{I}) \tag{7.4}$$

In general, neither the camera position nor the surface normals are known and—except in recordings under very controlled conditions—the nature and colour of the illumination source are rarely provided as meta-data with the video content—. Therefore, the illumination model defined in equation 7.3 needs to be somehow simplified.

In the proposed region segmentation technique we rely on two complementary illumination features—the angle between colour vectors and the albedo ratio—which are integrated in an original and effective way in the operation of Mean-Shift. Additionally—as a by-product of its application—the approach here defined generally improves BS operation, not only by correctly classifying moving cast shadows, but also by correcting wrong assignments of foreground pixels $\Omega_{B,F}(\mathbf{I})$ to the background set, and wrong assignments of background pixels $\Omega_{F,B}(\mathbf{I})$ to the foreground set.

Two simple illumination-related features.

Albedo ratio. In Nayar and Bolle [1996] the authors prove that, under strong assumptions, the albedo ratio is independent of the reflectance function and of the illumination spectrum. If we consider that the light source is white coloured and that the sensor response remains constant across the visible light spectrum, equation 7.3 becomes:

$$I = \varsigma \cdot e \cdot \rho \cdot R(\theta, \upsilon, \mathrm{n}) \tag{7.5}$$

, where dependence with the wavelength Λ has disappeared, ρ represents the integral of the reflectance function over the visible light spectrum, and $R(\theta, \upsilon, \mathrm{n})$ is the distribution of the reflected light for the particular wavelength of the incident light, hence discriminating between reflective power and reflectance distribution.

If we now consider a particular pixel in a small area surrounded by a smooth continuous surface, we can assume that θ, υ and n are approximately the same for every neighbouring pixel inside such area. According to this simplification we can define for two neighbouring pixels:

$$I_1 = k_1 \cdot \rho_1 \cdot R(\theta, \upsilon, \mathrm{n}), \quad I_2 = k_2 \cdot \rho_2 \cdot R(\theta, \upsilon, \mathrm{n}) \tag{7.6}$$

, where $k_1 = k_2 = k$ depend on the light source and on the sensor response.

From equation 7.6 it is derived that neighbouring pixels captured under the described conditions would show equal reflective power if they belonged to the same material but unequal if they belonged to different ones.

Note that this abstraction aims to provide a physical framework to the *obvious* assumption usually performed by region segmentation approaches: neighbouring pixels with similar brightness values are prone to be fused in a single region. The albedo ratio, defined as it follows, will be an indicator of this situation able to operate with independence of the brightness magnitude:

$$P = \frac{\rho_1}{\rho_2} = \frac{I_1}{I_2} \quad \text{or,} \quad P^* = \left| \frac{I_2 - I_1}{I_2 + I_1} \right| \tag{7.7}$$

, to avoid indetermination when $I_2 \approx 0$. As we are computing the reflectance ratio between neighbouring pixels, we can assume that all of them are illuminated with the same distribution emitted by the same sources. Hence, equation. 7.7 also holds for multiple illumination sources. All these expressions assume that pixel intensity has previously been gamma-compensated.

Angle between colour vectors

In the above model derivation, the term k has been defined as dependant on the camera sensor, the source spectrum and the source intensity. As claimed in Nadimi and Bhanu [2004], the sensor and spectrum can be considered the same for neighbouring pixels under certain conditions; albeit, source intensity can vary inside a reflectance-homogeneous region. If this is the case, a situation closely related to shadow and highlight presence, the model derived in equation 7.7 does not hold—as $k_1 \neq k_2$—.

When an object blocks a light source, the area behind the object in the trajectory defined by the light wave and the object becomes darker. This darkening can result in medium illuminated areas—penumbra—or poorly illuminated areas—umbra—depending on the relative position of the area with respect to the occluding object, the light source and the ambient illumination. Highlights represent a completely different effect. Specular surfaces reflect the incident light in a single spatial direction. If this direction is captured by the recording camera, the effect is the creation of scene areas which appearance is—partially or completely—occluded by the colour of the reflected light.

In these situations, pixels belonging to the same material but in different shady areas will not share similar albedo. To tackle this problem we propose a simple solution based on the use of the colour vectors angle as described in Dony and Wesolkowski [1999], which, although it does not respect the blueish characteristics of the shadow—darkening of the blue channel tends to be lower than in the other two channels—nor the strong directional responses of highlights, is able to efficiently handle moderate changes in illumination intensity if applied locally, as is our case,

without the requirement of training.

In particular, the technique assumes that, in a gamma-compensated RGB space, if a pixel with a colour vector $\mathbf{c}$ becomes under-illuminated, its modified colour vector can be expressed as $\mathbf{c}' = \alpha\mathbf{c}$—or $1/\alpha$ for moderate *highlighted* areas—. Therefore, both vectors share the same orientation. Note that the proportionality factor α is closely related with the light intensity component k in equation 7.6. This scheme has some problems derived from the structure of the RGB colour space (e.g., under low illumination conditions, colours are all almost proportional among them due to quantification), which have to be considered when using this descriptor. However, the use of the angle between RGB colour vectors would provide an intuitive and effective tool to assign moderately shadowed and highlighted pixels to background.

Using invariants to drive region extraction.

Mean-shift (MS) has been previously described in this document (see chapters 3 and 4). Let us here briefly review its conceptual application when used to segment images in order to provide a coherent explanation of the proposed system—further details can be found in the referenced chapters—. The objective of RS by MS is to find local extrema (peaks, modes) in the density distribution of a data set. For continuous distributions, MS just iteratively hill-climbs over the density distribution until it reaches a maximum. To provide robustness, MS works in a delimited part of the distribution. The window that encloses MS working area is defined by a kernel and the size of the window by the bandwidth of that kernel. This way, the technique avoids the influence of outliers in peaks estimation and, by shifting the window, is able to compute a set of peaks or modes that implicitly divide the data into a bandwidth-dependent number of clusters. These clusters are commonly fused in a post processing stage based on similarity criteria in order to avoid inaccuracies in the clustering owing to the bandwidth restriction.

MS is the best tool to support the combination of the proposed features for two main reasons:

- The method avoids the selection of fusion rules, which usually turns to be heuristic.
- There is no need to estimate the number of clusters for each frame and for each video.

The novelty of the proposed approach lies in the integration of the two presented features in the base operation of the MS algorithm: just the albedo ratio is used in the clustering phase and both features are used in the cluster fusion phase.

Bandwidth selection

The kernel bandwidth is a MS parameter that controls the criteria or restrictions to cluster pixels in the mode-seeking stage. Most MS-based segmentation approaches consider several pixel features (e.g., position, luminance, colour) and, for each, a similarity range. These jointly

define a multidimensional bandwidth. The use of this range implicitly assumes pixel-feature comparison via the Euclidean distance. We propose to combine pixel position, compared with the Euclidean distance, and pixel intensity, this compared via its ratio, hence inherently including the albedo ratio in the bandwidth selection.

In this line, we heuristically establish that a neighbourhood of a pixel is defined as the set of pixels that are spatially closer than 10 pixels ($h_{\mathbf{p}} = 10$) and present albedo ratios smaller than 0.01 ($h_{\Omega} = 0.01$). These parameters are set motivated by the assumptions made in Nayar and Bolle [1996] and in order to bypass the bandwidth selection process described in chapter 4—which, for computational reasons, cannot be applied in the analysis of video sequences—hence, simplifying the MS operation.

Mode fusion

The designed MS approach clusters regions attending to local estimations. This operation over-segments the scene in a large set of small regions, which are supposed to be reflectance-homogeneous in our case. According to a classical RS by MS technique, a second stage of the algorithm performs mode fusion, which is typically based on inter-mode similarity evaluation. This is commonly carried out over the same set of features used for grouping around their centroids—generally, RGB-colour medians—. In order to avoid shadow influence, in the proposed version of the algorithm we also include the colour vector angle in the fusion procedure.

The designed technique first searches for adjacent regions with RGB-colour median vectors whose cross product is close to 0—i.e., they are parallel or in this case collinear—. In particular, we evaluate the fusion of two adjacent regions if their cross-product is less than 0.01. As the albedo ratio restrictions should be conserved after mode-fusion, every pair of connected regions satisfying the angle restriction are further examined: first, the α proportionality factor between their respective centroid colour vectors is estimated by dividing them; then, this factor is used to correct the illumination intensity influence; finally, the albedo ratio is re-computed and the same similarity criteria applied in the clustering stage is applied to either merge or not the pair of connected regions. Figure 7.1 exemplifies the algorithm operation and sketches the proposed fusion scheme.

Expanding foreground detection results.

Segmented regions are assigned either to the foreground or to the background according to the number of foreground and background pixels that mostly contain, which is indicated by a simple input pixel level segmentation approach (García and Bescós [2008]). Through this method the motion information (static background objects, moving foreground objects) is expanded from correctly to incorrectly classified pixels according to region evidences by exploiting spatial constrains. Furthermore, the final region-level segmentation mask is used to update

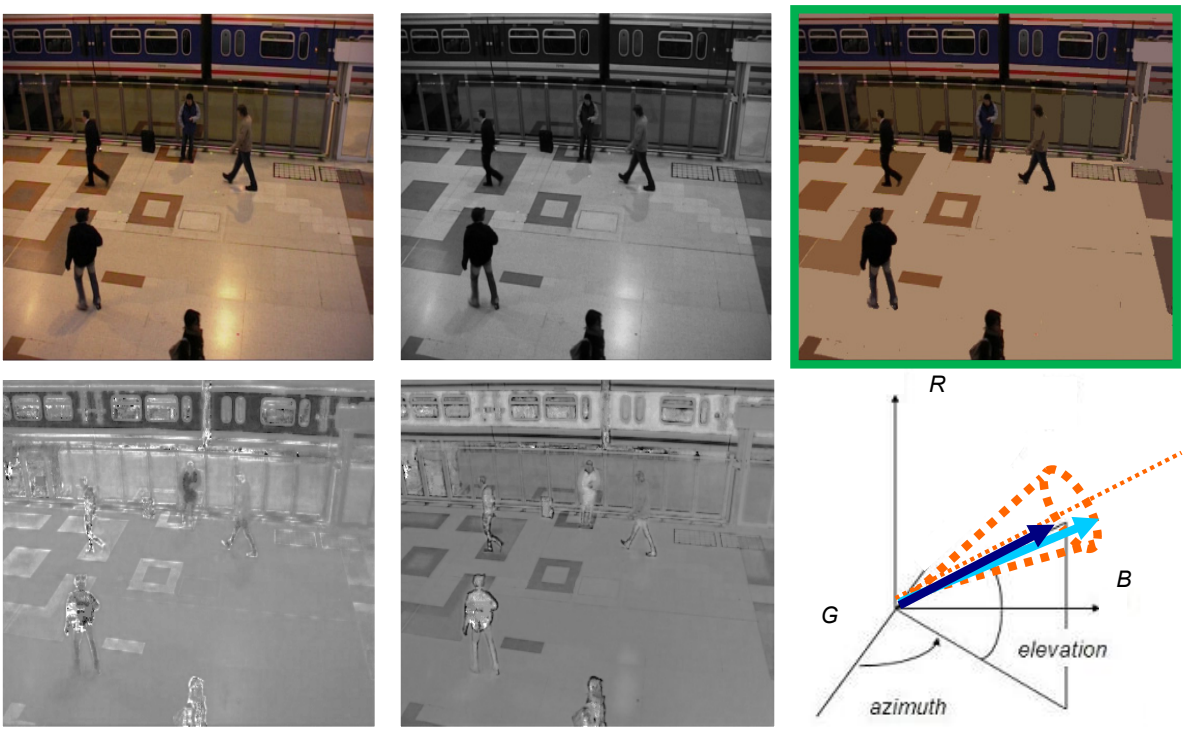


Fig. 7.1. Simplified features for reflectance-homogeneous region fusion. From left to right and top to bottom. First row: original RGB image, luminance, segmentation result after mode-fusion (with regions represented by their RGB-colour median vector). Second row: RGB-colour median vectors elevation, azimuth angle and sketch of the proposed fusion scheme. The fusion of the regions which RGB-colour median vectors are represented by the light and dark blue arrows is evaluated as these are close in the angular space. Note that colour vectors of pixels in the floor tend to share similar elevation and azimuth angles with independence of the illumination intensity that they reflect.

the background model at pixel level, hence, eliminating the temporal influence of—correctly reclassified—segmentation errors.

Experiments description.

We evaluate the proposed approach on selected sequences from the AVSS2007[1] dataset and the PETS2006[2] dataset. We want to express our gratitude to authors of (SanMiguel and Martínez [2009]) for lending us a manually annotated Ground-Truth (GT) of those sequences.

Results show the improvements achieved over three example videos: performance on shadows removal is qualitatively analysed in Figure 7.2—final masks in the bottom row—and quantitatively in Table 7.1.

[1] http://www.avss2007.org
[2] http://www.pets2006.net

Statistic	**Sequence**	PETS_S1_C3	PETS_S4_T5	AVSS07_AB
$\Omega_S^{GT}(\mathbf{I}) \in \Omega_{F*}(\mathbf{I})$ (%)	PM	63.39	34.22	87.38
	RM	7.19	3.64	3.22
$\Omega_S^{GT}(\mathbf{I}) \in \Omega_{B*}(\mathbf{I})$ (%)	PM	36.61	65.78	12.60
	RM	92.81	96.36	96.78
$1 - Sp$	PM	0.014	0.070	0.159
	RM	0.007	0.026	0.016
R (%)	PM	87.53	75.42	82.72
	RM	88.98	80.42	80.15
P_F (%)	PM	67.27	79.67	21.36
	RM	82.01	90.68	72.26
P_B (%)	PM	99.57	99.17	98.93
	RM	99.63	99.33	98.95

Table 7.1: Quantitative comparison of the invariant-to-illumination region-enhanced BS (RM) and the preliminary pixel-based BS(PM)—García and Bescós [2008]—. Global values per sequence.

As the GT contains annotations of the shadow $\Omega_S^{GT}(\mathbf{I})$, foreground $\Omega_F^{GT}(\mathbf{I})$ and background $\Omega_B^{GT}(\mathbf{I})$ pixels, Table 7.1 compares the refinement operation in two terms.

First, we compare the shadow correction by computing the fraction of shadow pixels in the GT, $\Omega_S^{GT}(\mathbf{I})$ that are assigned to the foreground, $\Omega_{F*}(\mathbf{I})$, or to the background, $\Omega_{B*}(\mathbf{I})$, in the partitions by the initial pixel-level segmentation mask and by the proposed region-based process. Note, that, for the pixel-level segmentation, the set of background pixels and the set of shadow-corrected background pixels is the same, $\Omega_{B*}(\mathbf{I}) \equiv \Omega_B(\mathbf{I})$, as no correction has been done. The same situation applies for the foreground set and the shadow-corrected set.

Second, several statistics at pixel level have been also computed on the foreground and background sets defined by the GT $\Omega_F^{GT}(\mathbf{I})$ and $\Omega_B^{GT}(\mathbf{I})$. These are extracted in order to measure the re-classification of highlighted pixels and small errors ($\Omega_{B,F}(\mathbf{I})$ and $\Omega_{F,B}(\mathbf{I})$) performed by the proposed refinement method. To this aim, we extract the following statistics: recall (R), 1-specificity ($1 - Sp$), foreground (P_F) and background (P_B) precision (see chapter 6 for definitions).

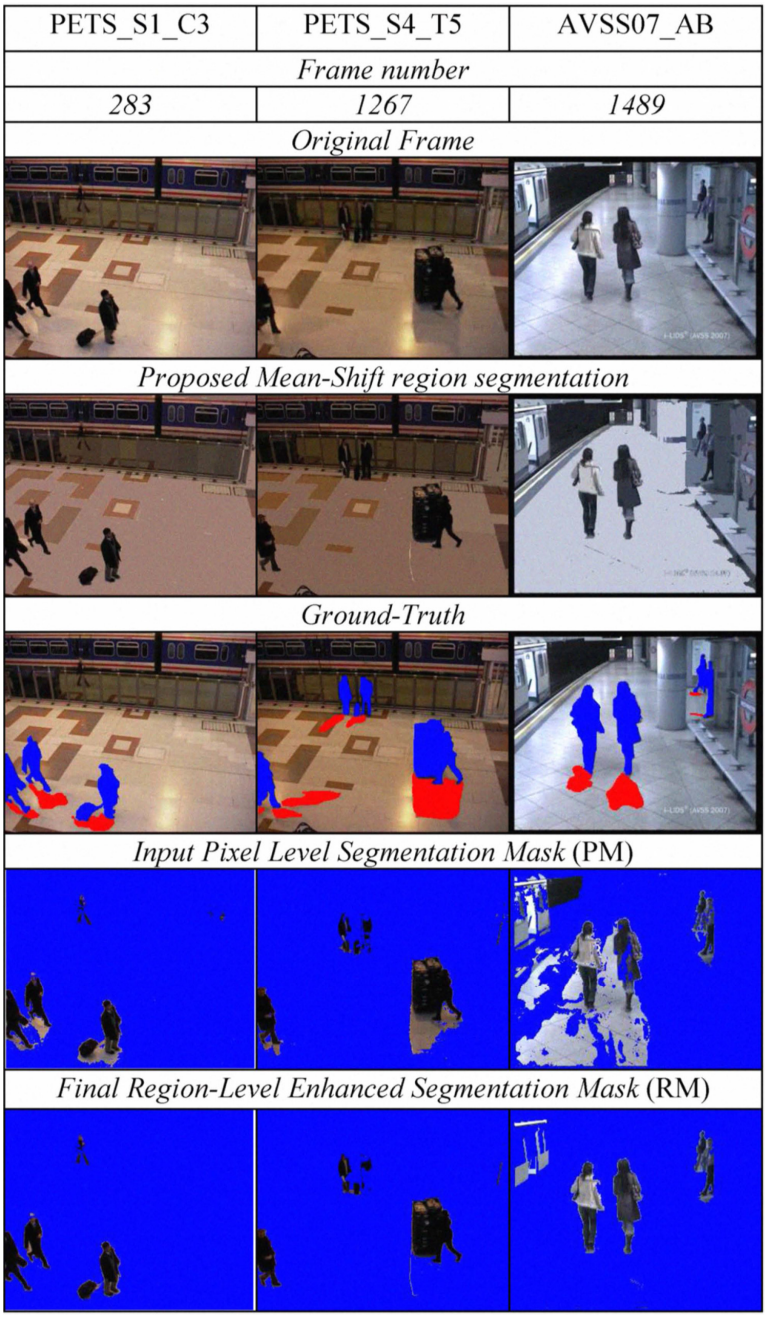


Fig. 7.2. Qualitative comparison of the invariant-to-illumination region-enhanced BS (RM) and the preliminary pixel-based BS (PM)—García and Bescós [2008]—at some example frames. See how region continuity in the floor solves the cast shadow problems.

Results discussion and approach limitations

In the light of the quantitative and qualitative results, it can be stated that, on the analysed sequences, the use of a simple, but effective, RS scheme based on illumination invariants improves the performance of a pixel-based BS algorithm in several terms. First, the region-enhanced version of the algorithm adequately reassigned most of the shadow samples to the background—see second and third rows of Table 7.1 and the last two rows of Figure 7.2—. Additionally, according to Table 7.1, the accuracy in the classification of foreground samples is severely increased—an average improvement of a 91 %—without significantly altering the classification recall—an average increase of a 2 %— Finally, the system sensitivity is also moderately improved.

The designed approach might be used as an alternative to classical shadow-removal post-processing algorithms but assuming several limitations. Obtained regions are homogeneous just respect to the albedo, hence, intensity patterns which may shape a continuous area according to any other feature might be over-partitioned—, i.e. a tiled floor or a brick wall are homogeneous in texture, but not necessarily in albedo (see examples in chapter 4)—. Furthermore, foreground evidences are here initialized by a pixel-based method; hence, regions which pixels are not mostly assigned to foreground would produce a counterproductive effect in the mask refinements. This effect would be particularly noticeable in camouflaged areas of the frame. These areas would tend to be wrongly assigned to the background by the BS approach. With the proposed refinement method, this wrong assignation wont be corrected and even could be extended to similar foreground areas in their neighbourhood, with independence of the background modelled for these areas. In other words, severe failures of the BS approach—foreground areas with less than half of their pixels identified as foreground—might be extended by the region-based-refinement approach.

A further development of the approach is therefore required, either by integrating alternative segmentation methods (as those described in chapters 3, 4 and 5), by incorporating spatial information by alternative schemes or by introducing additional processing on the regions—this is inspected by the next solution—. Nevertheless, the results of the proposed approach respect to the BS solution evaluated suggest that introducing spatial information may be useful for improving the operation of BS methods.

7.2 Case of example 2: A multi-layer region-based model for background subtraction

The aim of the example here described is the proposal of a region-based model that—by operating together with a pixel-level BS method—is able to yield tightened-to-objects segmentation masks in complex scenarios without requiring further post-processing. Specifically, the proposed approach starts from a RS of each frame obtained through the method described in the previous

section 7.1. So-obtained regions are used to build and update a multilayer background model and a foreground model. Regions in the models are characterised by a time varying covariance matrix which encloses a set of spectral and spatial features. The evolution of the covariance matrix describing each modelled region is used to discriminate between foreground and background regions. After region discrimination, a simple feedback scheme exports segmentation results to the pixel-level BS algorithm.

The approach presented in this section is—to our knowledge—the first defining a region-based model of the background. This fact requires the adaptation of classical pixel-based tools to the region-based scope. The next paragraphs are devoted to define these processes and present preliminary results of the so-designed region-driven BS method.

Region characterization

The algorithm operation relies on the previously described illumination-blind RS (see section 7.1). Each frame $\mathbf{I}_t$ is partitioned into regions by applying the MS technique on the albedo ratio; adjacent regions are then fused if their normalised colour vectors are collinear and their gamma-corrected luminance channel respects the albedo condition. These regions are then characterized as it follows.

Let $\Omega_{t,k}$ be an instance of a region k at frame $\mathbf{I}_t$, and let $\boldsymbol{\varphi}_{t,k}$ be the feature vector that characterizes it:

$$\boldsymbol{\varphi}_{t,k} = \{\varphi^1_{t,k}, \varphi^2_{t,k}, ...\varphi^N_{t,k}\} \tag{7.8}$$

For characterisation purposes we have selected a set of simple spectral features for each region: the RGB-colour median of the region (a three-component vector); the region area; the number of pixels that the region encloses (one value); and the RGB-colour median angles respect to two reference vectors (two values), which, combined, provide an additional feature to characterise regions with light intensity varying frame to frame. Furthermore, we have included a set of location features: the RGB-colour median angles respect to 8-connected adjacent neighbours $\mathcal{N}_8(\Omega_{t,k})$ of the region. Note that—as our aim is to provide a characterisation robust to region rotations—we have just selected a neighbouring region in each of the eight orientations—discarding the rest of adjacent regions if any—. Furthermore, regions which are adjacent in one or more of these eight orientations to the image boundaries are declared as adjacent to themselves—thus, characterised with *0 rad* angles in these directions—. Therefore the so-designed description vector is always composed of $N = 3 + 1 + 2 + 8 = 14$ features. However, as alternative characterization schemes can be naturally incorporated to the model we keep the generic symbol N from here on.

We obtain the $N \times N$ temporal covariance matrix of the features that characterize the T

last instances of region k by:

$$C_{t,k}(i,j) = \frac{1}{T} \sum_{\tau=t-T}^{t} (\varphi^i_{\tau,k} - \mu^i_{t,k})(\varphi^j_{\tau,k} - \mu^j_{t,k}) \tag{7.9}$$

, where $t \geq T$, and:

$$\boldsymbol{\mu}_{t,k} = \{\mu^1_{t,k}, \mu^2_{t,k}, ...\mu^N_{t,k}\},\ \mu^j_{t,k} = \frac{1}{T} \sum_{\tau=t-T}^{t} \varphi^i_{\tau,k} \tag{7.10}$$

, is the arithmetic mean vector in the temporal period $[t-T, t]$.

Given a potential instance of region k at frame $\mathbf{I}_{t+1}$, we define the cost of updating the temporal covariance matrix as the distance between $C_{t,k}$ and $C_{t+1,k}$, computed with the measure proposed in Förstner and Moonen [2003] and later used—for different purposes—in Tuzel et al. [2006, 2008]; Wang et al. [2012]:

$$D_{t+1,k}(C_{t,k}, C_{t+1,k}) = \sqrt{\sum_{j=1}^{N} \ln^2(\lambda_j(C_{t,k}, C_{t+1,k}))} \tag{7.11}$$

, where $\lambda_j(C_{t,k}, C_{t+1,k})$ are the generalised eigenvalues of the covariance matrices, obtained by solving:

$$\lambda_j C_{t,k} \boldsymbol{\nu}_j - C_{t+1,k} \boldsymbol{\nu}_j = 0,\ j = 1...N \tag{7.12}$$

, being $\boldsymbol{\nu}_j$ the generalised eigenvector associated to λ_j.

Generalized eigenvalues computation requires matrices to be positive definite, i.e. all of their principal leading minors have to be positive, which might not be the case if a region feature does not vary frame to frame. Hence, we eliminate in both matrices the rows and columns which correspond to non positive principal leading minors in any of them. In case every principal leading minor in one or both of the matrices is not positive, noise is added to their diagonal until at least one becomes positive for both matrices.

The cost of updating a covariance matrix provides a robust measure of similarity between previous region instances and every new region instance. For matching purposes and according to the nature of the camera noise, we further propose to model the evolution of the cost of updating a covariance matrix, with a single Gaussian updated with a classical Running Average scheme:

$$\mu_{D_{t+1,k}} = \alpha \mu_{D_{t,k}} + (1-\alpha) D_{t+1,k}(C_{t,k}, C_{t+1,k}) \tag{7.13}$$

and:

$$\sigma_{D_{t+1,k}} = \alpha\sigma_{D_{t,k}} + (1-\alpha)\left|\mu_{D_{t+1,k}} - D_{t+1,k}(C_{t,k}, C_{t+1,k})\right| \tag{7.14}$$

However, as the temporal evolution of the metric defined in 7.11 is not ensured to be Gaussian, alternative modelling schemes need to be inspected in the future work.

Region comparison

Following the described procedure, for every region k at frame $\mathbf{I}_t$ we account for a set of K modelled regions, $\Omega_{t,k}$, $k = 1...K$—the number of modelled regions and their spatial arrangement are described later in this section—. Each of these modelled regions are characterized via its covariance matrix, $C_{t,k}$.

For a new frame $\mathbf{I}_{t+1}$, in order to match a given region $\Omega_{t+1,\kappa}$, to the modelled ones, we calculate the set of $N \times N$ temporal covariance matrices, $C_{t+1,k}$, resulting from considering that this given region is a new instance of each k^{th} modelled region. We then compute the corresponding set of updating costs, $\{D_{t+1,k}\}$, as defined in equation 7.11. Next, the coherency of every k^{th} updating cost with the Gaussian-modelled cost evolution of the k^{th} region is checked. This results in the definition of a set of positive matches:

$$\Upsilon = \{k : |D_{t+1,k}(C_{t,k}, C_{t+1,k})| \leq \mu_{D_{t,k}} + 2\sigma_{D_{t+1,k}}\} \tag{7.15}$$

Finally, from this set of positive matches—if any—we select the optimal as the one with the minimum associated covariance distance:

$$\dot{k} = \underset{k\in\Upsilon}{arg\,min}(|D_{t+1,k}(C_{t,k}, C_{t+1,k})|) \tag{7.16}$$

Region-oriented modelling

The proposed background model exports pixel-level schemes to regions. As pixel-level BS techniques try to model the different values of each single pixel along the video to handle dynamic background situations, our aim is to model the different *variations* that each background region can undergo. This is achieved via a multilayer background model (see Figure 7.3): static regions—those which appearance does not vary in the video—tend to be modelled by the first layers in the model—they appear as black or empty in successive layers—, whereas each *variation* of a dynamic region is modelled by successive layers. Dynamic regions are mainly caused by unresolved illumination effects, temporal instability of the RS or background dynamism.

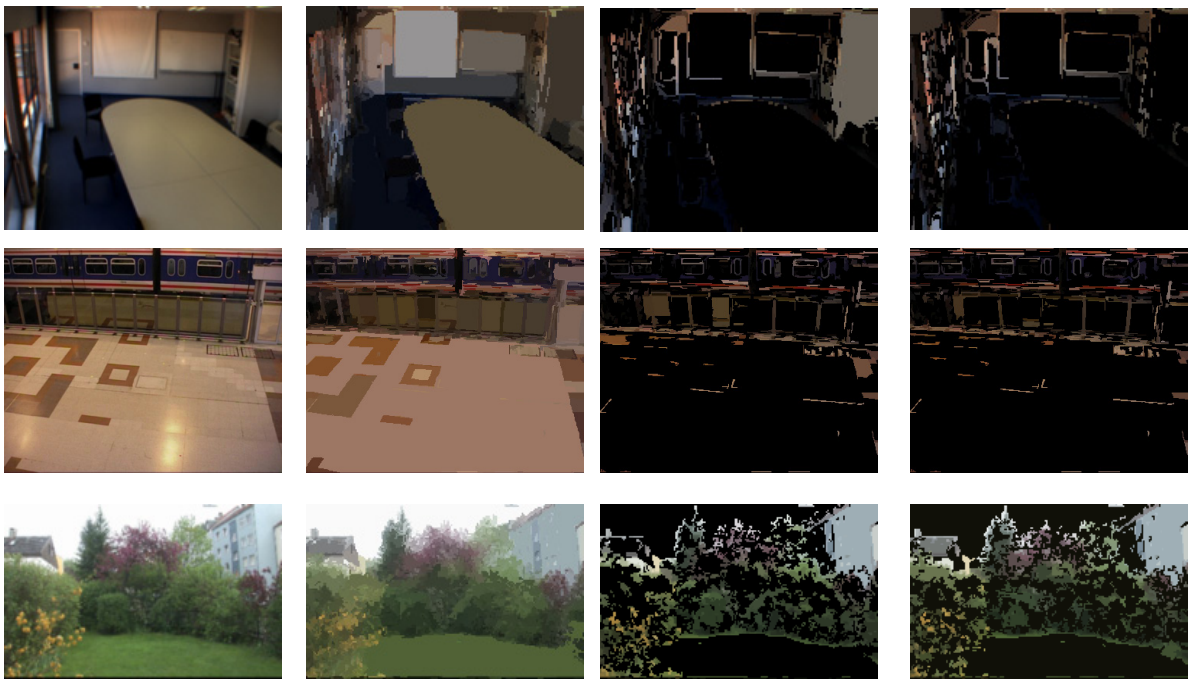


Fig. 7.3. To left. A given video frame. Next three columns: the three first layers of the region-based background model. In the first two rows *variations* are caused by unresolved illumination effects and RS instability. In the third row the principal cause of *variation* is background dynamism.

Model initialisation

The construction of the covariance matrix, as defined in equation 7.9, requires T region instances. Hence, region matching for the first T instances demands alternative matching schemes. In this stage, which we consider the model initialisation stage of our algorithm, region similarity is measured via Euclidean distance between feature vectors. This might harm region characterization, as selected features may be highly correlated. In order to reduce the influence of this problem, initialization time has been selected relatively small ($T = 10$).

Model maintenance and foreground detection

For every incoming frame, in parallel to its segmentation into regions, we perform a basic pixel-level segmentation—we use García and Bescós [2008], as in the previous case of example—to obtain an initial segmentation mask. This mask is used to pre-classify input regions. Regions that only overlap with the background mask are considered *Confirmed Background Regions* (CBRs) and regions that contain at least one foreground pixel are considered *Potential Foreground Regions* (PFRs).

A CBR updates the region that best matches in the background model: the search is performed layer by layer, testing with regions overlapping a circular area of radio r and decided according to equation 7.16. If no match is found, a new region is initialized in the first empty or black layer in its position.

A PFR undergoes the same matching process against the background model. If it results in a match to a background region, the PFR is re-considered as a CBR and updates the matched background region. Otherwise, the PFR is labelled as a confirmed foreground region (CFR) and it is used to build and maintain a foreground model.

The operation on the foreground model is equivalent to that of the background. The main differences lie in the mono-layer nature of the foreground model and in the searching circular area for matching, which should be bigger (we use $3r$) to account for moving objects displacements. Even though the advantages of this foreground modelling are not being fully exploited yet, the potential use of the foreground model to track moving objects is evident—see appendix B for an example of potential benefits of using a foreground model—. Finally, the obtained region level segmentation mask is used to update the pixel-level background model maintained by the pixel-level segmentation algorithm, therefore, theoretically improving its operation by avoiding the updating of correctly reclassified areas.

Experiments description

This section presents an initial quantitative and qualitative evaluation of the proposed algorithm. We compare our method with a State of Art (SoA) algorithm Li et al. [2004], in which results

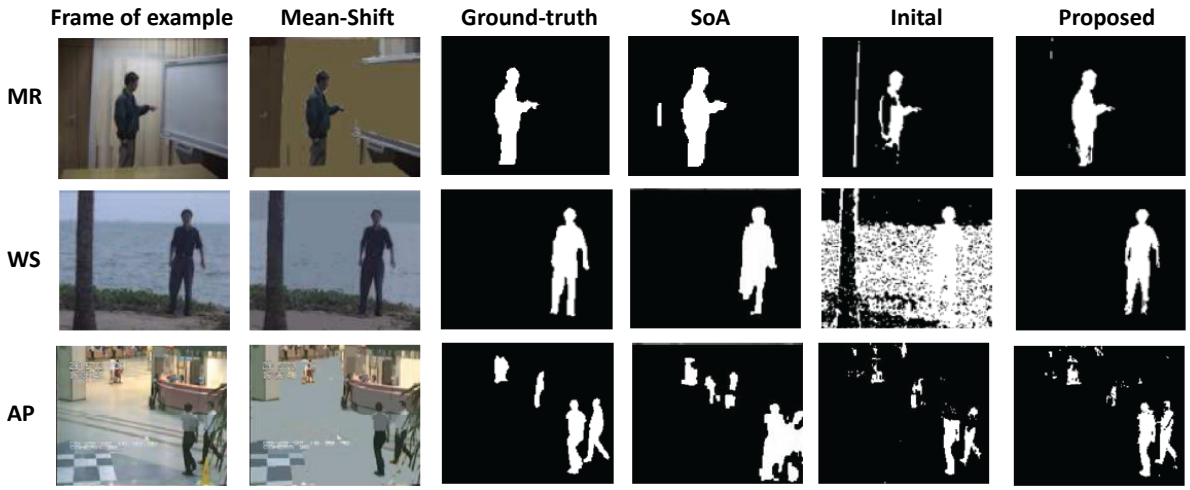


Fig. 7.4. Qualitative comparison of the region-based multilayer BS with: a SoA approach—Li et al. [2004]—and the preliminary pixel-based BS (Initial)—García and Bescós [2008]—at some example frames of the Meeting room—MR, first row—, the water surface—WS, second row—and the Airport—AP, third row—sequences. See how the algorithm is able to yield tightened-to-object masks and handle background dynamism.

are extracted from a set of videos associated to a manually annotated ground-truth[3]. From this set we have tested three videos, named Meeting Room where background contains a periodically moving curtain, Water Surface where most of the background is moving water and Airport, a scene with illumination artefacts on the floor. Qualitative results for these three videos are presented in Figure 7.4, one per row. The first column includes a sample frame and the second its mean-shift segmentation. Next columns present, for such frame, the manual ground-truth, the SoA algorithm result (Li et al. [2004]), the proposed initial segmentation mask (García and Bescós [2008]), and the finally achieved segmentation mask.

Quantitative results have been computed for every frame with available ground-truth (not all of them). To faithfully compare with Li et al. [2004], we have followed the same similarity measure proposed there; the overlapping between background and ground-truth masks: $O(\Omega_F^{GT}(\mathbf{I}), \Omega_{F*}(\mathbf{I}))$ (see equation 3.4).

Results discussion and approach limitations

Results are comparable to those reported by the SoA algorithm. However, first, we do not perform any kind of post-processing—which dramatically increases the mask quality in most segmentation approaches—, hence avoiding object-size dependent morphological operations. Second, detected inaccuracies in the manually generated ground truth—observe that the pro-

[3] http://perception.i2r.a-star.edu.sg/bk_model/bk_index.html

Statistic	Sequence	Meeting room	Water surface	Airport
$O(\Omega_F^{GT}(\mathbf{I}), \Omega_{F*}(\mathbf{I}))$	SoA	0.911	0.851	0.508
	Initial	0.300	0.156	0.493
	Proposed	0.899	0.822	0.494

Table 7.2: Quantitative comparison of the region-based multilayer BS with: a SoA approach—Li et al. [2004]—and the preliminary pixel-based BS (Initial)—García and Bescós [2008]—. Global statics per sequence in terms of foreground mask overlapping.

posed algorithm's mask seems to be tighter to the real objects than the ground-truth—may bias negatively our results. Nevertheless, the algorithm is able to improve a given input segmentation mask along the video, and the results presented here, yet too incomplete, are promising in our opinion.

The limitations of the proposed approach are mainly related with the inability of regions to properly guide the BS process without the support of a pixel-level BS method. The RS decisions are made per frame, ignoring previous and posterior results and hence, the regions are unstable from frame to frame, making their robust matching very problematic. Therefore, as the method depends on a previous BS approach, it inherits part of the defects—but also part of the advantages of these methods—. Recent schemes proposing RS in the spatio-temporal domain Liang et al. [2014]—i.e. relating regions along the video—may provide a solution to this problem. However, the operation of these approaches is still restricted to the analysis of short videos, due to a constrain in the number of traceable regions—a fixed number of regions is usually imposed—.

7.3 Chapter conclusions.

The first case of example in section 7.1 presented a simple method to incorporate illumination invariants in the region segmentation process. These invariants were derived from a physical illumination model under strong, yet realistic, assumptions. The potential benefits of the designed method were illustrated by its use to expand motion evidences in a BS approach. The combined approach presents remarkable benefits over the non-enhanced BS algorithm. Despite the fact that its use in combination with superior BS algorithms might not provide the same level of improvement as well as its discussed potential limitations, the designed system may be understood as an alternative to pixel-based shadow-removal post-processing methods. However, it is unable to face background multi-modality.

Regarding the second case of example (see section 7.2), it presented a pixel-level segmentation

approach which, thanks to the use of a complementary region-level analysis, was partially robust to illumination artefacts and robust to small segmentation artefacts, which are instead traditionally solved by post-processing stages. The proposed covariance-based region modelling and matching provided a robust and feature-scalable solution while accounting for plausible feature-correlation. The main novelty of the proposed approach relies on the use of an eigenvector-based comparison measure in a region-oriented BS model. Based on these region-management contributions, the presented region-based segmentation framework achieves accurate results in complex situations, while allowing for promising tracking possibilities. However, the evaluation of the proposed scheme as a complement of a BS approach of a higher quality is also required to check if the improvement percentages are of a similar order.

In spite of their promising results, a proper evaluation of the proposed approaches in a bigger and more complete dataset—the obvious choice should be the change detection dataset (Wang et al. [2014b])—is still required.

Part IV

Part IV. Regions for description constraining

Contents

This part deals with the local description of image points in two and three dimensional scenarios.

In chapter 8 we face the task of severe occluded object identification in Kinect-like scenarios. We assume that objects have been previously segregated and propose a solution to identify the objects by having a small visible evidence of their appearance. To this aim, we rely on the use of regions to generate a multi-coarse RS in order to cope with a varied set of potential occlusions. These regions are used to spatially-constrain the descriptions of two-dimensional and three-dimensional state-of-the-art point-of-interest descriptors. The method requires a very small and low-varied set of training data and generalise training knowledge by means of a neural model. The contents of this chapter are completed with Appendix C.

In chapter 9 addresses the matching of image points across scene views captured by widely separated cameras. To this aim we propose to define diffuse supports for descriptions. The size of these supports is automatically defined by evaluating the sparsity of in-support inter-pixel resemblances. So-obtained supports are projectively transformed under a set of geometrical transformations. The nature and number of these transformations are constrained by the scene calibration. Appearance transformations of the anchor support are also considered by locally adapting the description features through a lineal weighted scheme. The designed approach is tested on scenarios on which state-of-the-art solutions do not perform correctly.

"To observe attentively is to remember distinctly."

Edgar Allan Poe. (The Murders in the Rue Morgue, 1841).

Chapter 8

Severe-occluded object identification via region-based descriptions

This chapter presents a region-based strategy for part-based object identification. Identification stands for the capability of recognising an object instance—e.g. a particular cup—with independence of the external factors that affect its captured image—including light variations, capture point-of-view or occlusions—. Note that this problem is different—and much more simple—than the recognition problem, which instead, aims to identify object classes—i.e. cups—.

Starting only from colour images and depth estimations—i.e. not requiring complete 3-dimensional models—, we focus on the identification of trained objects in severe-occlusion scenarios; hence, we can call this problem re-identification. Two main stages shape this task: object segregation from the scene and object identification. To face this scenario, we assume that objects have been preliminarily segregated from the scene, so, we only intend to identify them. Note, that whereas this problem seems to be a simple one, strong changes of appearance—due to one or several of the aforementioned factors or to the object nature, e.g. deformable objects—substantially increase the problem complexity.

The proposed algorithm starts from the segregated assumption and follows by splitting segregated objects in successively coarser region-partitions; with each region representing a part of the object from which it was extracted. For the characterization of these parts, two new region-driven descriptors are proposed: R-DAISY and R-SHOT. The former encapsulates luminance and depth information inside a region with a DAISY-like [Tola et al., 2010] organisation, whereas the latter arranges surrounding normals and colours of three dimensional singular-points in a SHOT-like [Salti et al., 2014] scheme. Their novelty relies on the use of a size-and-shape-variable description support which is automatically defined by the object part itself.

So-obtained descriptions are self-organised in a single neural structure by an unsupervised learning process. This structure allows to automatically discover relations between the object's

parts. The information of an object part is then encoded in a distributed fashion by a set of *signatures*, each corresponding to the response of the neural structure to an object part description. Experimental results show that the approach achieves promising results in the identification of severe-occluded objects relying only on Kinect-captured data and using a very small set of training instances—2-to-8 short-varied Kinect-captured views per object to identify—.

The chapter starts by a review of existing approaches in which the state-of-the-art is related with classical human perception theories (section 8.1). This review helps to motivate the proposed approach and to introduce the main ideas used along the chapter (section 8.2). The chapter continues with an overview of the proposed approach in section 8.3. Characterisation and knowledge modelling stages of the algorithm are described in sections 8.4 and 8.5 whereas identification is described in section 8.6. The performance of the designed approach is evaluated in section 8.7 . Finally, section 8.8 concludes the chapter.

8.1 A review of existing approaches with a connection to human perception

Humans are able to identify an object in a wide range of sizes and points of view. In order to achieve invariance respect to these factors, the visual information is supposed to be projected from the retinotopic organisation over a cerebral area, so that projections from various retinotopic locations converge over the same invariant area. Despite the fact that when we identify an object we are quite aware of its relative size, position and rotation, the identification process is highly robust to changes in these aspects. There are two main information-codification theories trying to explain this fact ([Goldstein, 2002; Hoffman and Logothetis, 2009]) object centred and viewer centred.

The **object centred** perspective is explained from a psychophysical point of view. Object information is assumed to undergo a set of transformations until it matches a single stored 3D template. We can find a common path among the studies in this line. In a first stage some sort of features are extracted (depending on the theory, they range from simple shapes, lines, edges or salient areas to volumetric primitives); then, these are sorted and combined. In the final stage the object is identified by formulating a query to our stored knowledge (Marr [1982]; Treisman [1993]; Biederman [1987]). In machine vision applications inspired by these theories, the connection is particularly noticeable in the object characterisation stage. The use of different types of gradient-based descriptions—as in the first layers of the Marr computational theory ([Marr, 1982])—has been widely reported in the literature, either looking for salient and invariant-to-appearance points or regions ([Mikolajczyk et al., 2005]), or organised in local histograms—e.g. via HoG ([Kinnunen et al., 2012])—, which is also the feature used to build the well know deformable part models ([Felzenszwalb et al., 2010a]). Alternatively, object description via 2D and 3D primitives

was proposed in [Hu and Zhu, 2010], whereas descriptions via shape and Textons were the core of [Khan et al., 2012; Mori et al., 2005] and [Zhu et al., 2005] respectively. These essence-based descriptions are closely related to the elementary features described in the Feature Integration Theory ([Treisman, 1993]) in the case of schemes based on primitives; and closely related to the Biederman's axon based studies ([Biederman, 1987]) in the case of Texton-based schemes..

The **viewer centred** perspective assumes that knowledge increases with experience. Having observed a *representative* number of views of an object, new observations may be easily matched with the previous observed set ([Hoffman and Logothetis, 2009]). Almost every artificial training system is coherent with this assumption. However, how many views are required to reasonably succeed in the identification of objects from new points of view? State-of-the-art methods require either the tedious collection and annotation of large data corpora to learn object models, or the use of both detailed and parametrizable 3D object models, as the well known CAD models. The use of features derived from these models has been proven to be successful in the identification of objects in complex scenarios ([Hinterstoisser et al., 2012]). However, they require the existence or pre-construction of the CAD model, a requirement that hinders its scalability as well as its use in non-expert oriented applications. Furthermore, their operation is usually limited by the sampled model's orientations used for training, and, when used to develop holistic models, by the presence of occlusion. In this sense, region-based approaches—already well established in 2D scenarios: [Arbelaez et al., 2012; Felzenszwalb et al., 2010a; Frome et al., 2004; Felzenszwalb and Huttenlocher, 2005]—, as pictorial ones, are supposed to operate well in the eventuality of partial occlusions and, if applied adequately, may be more robust to unobserved variations—hence untrained—of the object's appearance than holistic approaches.

In this study, we face the problem of identifying arbitrarily textured 3D objects captured with a Kinect camera, in the absence of a 3D model nor of a huge training data-set for such objects. Objects models were trained using just a very small number of instances—in our experiments no more than 8 instances per object were used—are then used to identify new instances of these objects when these are captured from new points of view and/or under different illumination conditions.

Our aim is to identify objects when they are severely occluded by other objects, which prevents from using templates and holistic models. Most of the existing works facing this problem select or design a set of ideally robust and discriminative features, then used to determine correspondences between an object instance and the modelled objects. In this line, it is worth to read the works dealing with descriptions based on: point signatures ([Chua and Jarvis, 1997]), spin images ([Johnson and Hebert, 1999]), spherical spin images ([Ruiz-Correa et al., 2001]) and local surface patches ([Chen and Bhanu, 2007]). Alternatively, more recent studies explicitly devoted to object identification with the Kinect device have been also proposed ([Tombari and Di Stefano, 2010; Lai et al., 2011a; Bo et al., 2011; Mian et al., 2010; Lai et al., 2011c]). Some of them export

classical two dimensional object descriptions to $RGB-D$ situations (e.g. singular-points in [Mian et al., 2010] or signatures of histograms in [Salti et al., 2014]). Moreover, in order to provide robustness to occlusions, techniques based on Hough voting for both non deformable and deformable objects have been also proposed ([Tombari and Di Stefano, 2010]). Regarding the learning strategies, linear and non-linear support vector machines ([Lai et al., 2011b]) are usually preferred; however, Hierarchical Kernel descriptions ([Bo et al., 2011]) and tree-based approaches ([Gavrila and Philomin, 1999]) have been also used with relative success.

8.2 Main idea and motivation

It should be made clear that this study in no way intends to partially replicate how the human visual system works; instead, it aims to *mimic* some operation mechanisms that are supposed to take place on it, as suggested by research results in the field of visual perception in highly-developed visual systems (in particular, this study is highly inspired by [Goldstein, 2002]). In our opinion, the existing computer vision approaches to object identification in *complex* scenarios are constrained not only by the features or the metrics used to model and compare object instances, but also by their operational paths and strategies. In essence, this study is mainly motivated by three premises or targets:

1. Define an object model which can be trained with a very small number of samples.
2. Confront object identification under severe occlusion situations.
3. Provide a distributed modelling approach which automatically arranges training evidences.

Training with a few samples

A critical component of vision is the creation of visual entities, that is, representations of surfaces and objects that do not change the perceived scene but change which parts we see as belonging to other objects and how these are arranged in depth. Humans learn object appearances by combining examples with their knowledge of the behaviour of the visible world—i.e. their expertise ([Wong et al., 2012])—.

The reader would not usually require a big amount of examples to re-identify an object when it is rotated or when it appears in a scenario different to that in the *learned* examples (few-shot learning [Rohrbach et al., 2013]). This is agreed to be achieved by the feature management mechanisms used to perceive objects by the ventral path ([Goodale and Milner, 1992]).

Back to the computer vision world, template-matching approaches train the object model with thousands of templates usually extracted from a CAD-model ([Hinterstoisser et al., 2012]).

Robustness to object rotation and scale change is in these cases obtained by quantifying the potential appearances of an object observed from a set of plausible points of view.

One of our targets is to model objects with a low number of training samples. To cope with object rotation, we propose to provide robustness in the description itself, via the use of a local reference frame ([Salti et al., 2011]), i.e. locally aligning descriptions with the object—or with a particular part of it—so that these are independent of the point of view at which the object is captured. This is not a novel approach: [Salti et al., 2014] have recently compiled reported studies in the design of such sort of descriptions. Nonetheless, in order to recover—at least putative—matchings between local referenced descriptions, the reference for the local alignment should be stable for different views of the same object; that is, the reference should be stable to point-of-view changes. To cope with scale changes, the scale-space theory [Lindeberg, 1993] establishes a well-founded mathematical framework to discover singular points of an object view that are recoverable to some degree from a moderate affine-transformed view of the object. In fact, this theory establishes the basis of classical point-of-interest detectors including the well-known SIFT points[Lowe, 2004].

Handling severe occlusion situations

When an object is occluded, only part of it—maybe down to 10%—is visible. A system aiming to identify the object in these scenarios should adapt its trained knowledge—observed samples, which in order to maximise the amount of training data are generally extracted from holistic examples—to an unpredictable occluding situation which always results in incomplete instances of the target object. The extent of these incomplete instances is visually defined by both the real contours of the object and the occluding contours of the interfering objects; hence, contours and holistic templates might not be a reliable cue for characterisation. From our perspective, there are two main ways of facing this situation.

The first one consists of fitting an holistic model to the visible and non-visible parts of the occluded instance, assuming that the not-visible part of the object remains unaltered. The likelihood of the instance being the tested object can be obtained by measuring the similarity between the instance's visible part and its corresponding part in the model ([Hinterstoisser et al., 2012]). This top-down approach—strongly linked with the Gestalt's principle of continuity ([Spelke, 1990])—may fail if the initial holistic fitting is inaccurate or if the visible parts are insufficient to establish a reliable correspondence to an specific part of the model.

The second one, a bottom-up alternative, consists of considering the occlusion in the learning process, i.e., dividing the object in its *semantic* parts and training each part independently. The instance may be then identified, at least partially, by integrating the likelihood of each identified part. A system driven by this philosophy should operate better in situations where only a small part of the object is visible, which is our objective. Advantages of using this part-based

modelling approach for the identification of cars at different 3D view-points were illustrated in [Sun et al., 2009].

Distributed model encoding

Humans have limited resources available for storing knowledge ([Gauthier, 2000]), which disregards the theories associated with the coding specificity, that is, the existence of devoted neurons—activated by a particular complex stimulus (e.g. a particular object or face)—. However, many of the existing studies on artificial object identification follow precisely this path: models are trained for specific objects and only the response of an object instance to a specific one among such models is considered for the identification. On the contrary, distributed encoding is the representation of a specific stimulus by an activation pattern distributed among several neurons. This scheme allows to represent a big number of stimuli with a smaller set of neurons.

In essence, the prevalence of one scheme over the other is assumed to be dependent of the nature of the stimuli: simple stimuli, like motion or contour orientation, seems to be perceived with specific stimuli-devoted neurons ([Newsome et al., 1989]); whereas complex stimuli, as faces or objects, require the combined activation and inhibition of sets of neurons—as it was suggested in the monkey-based study of [Abbott et al., 1996]—. Many experimental results in the area of cognitive perception have shown ([Deadwyler and Hampson, 1995; Eichenbaum, 1993]) that the independent firing of neurons is not the most significant factor for encoding information. The coordinated activity of assemblies of neurons relates to a functional organization of such assemblies, hence implying that assembled neurons fire together to common stimuli.

In our opinion, in machine-vision applications, the selection of one scheme or the other reflects in either designing and training a specific model for each considered object, and then testing each model against an object instance; or constructing a single common model for all considered objects, and then classifying an object instance by measuring its response to this model.

Approach statements

In the light of these reflections, and in order to fulfil the initial premises, our proposal presents a knowledge-model for object identification that: 1) describes an object via locally referenced descriptions extracted around its singular-points—for comparison, a description based on depth and colour distributions is also presented—; 2) encapsulates these descriptions in a set of object's parts extracted at different coarseness levels; and 3) self-organises, through an unsupervised learning process, the so-learnt knowledge on a single neural-oriented model in which responses to similar stimuli concentrate on the same neural area.

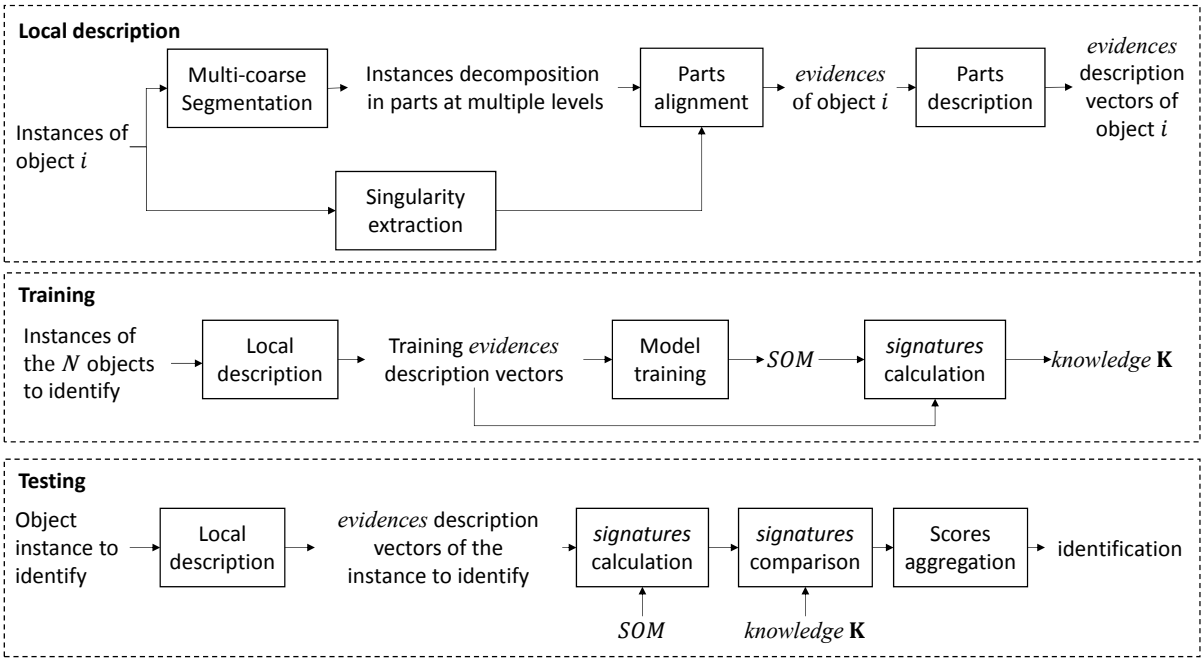


Fig. 8.1. Flowchart of the identification method. See graphical representations of each stage in Figures 8.2 and 8.3 and text for details.

The likelihood of an object instance being a modelled object is designed to account for both the *tuning* to the neurons modelling such object and the *inhibition* to the rest of the neurons, then following a distributed encoding philosophy. To provide enough experimental evidence for the results to be significant, a highly complex scenario should be faced whereas, in order to enhance experiments plausibility, robust descriptions need to be created.

In order for this scheme to succeed, three key design factors need to be addressed:

1. Robust object characterisation: the features that describe an object's part should be robust to appearance variations and view-point changes.

2. Stable object partition: object's parts should be stable and representative, i.e. we aim to obtain similar partitions in the training and testing stages.

3. Flexible object modelling: the knowledge-storage scheme should automatically group object's parts which descriptions are similar, hence defining common spaces of identification.

The rest of the chapter aims to describe the solutions proposed for these three design factors.

8.3 Approach overview

In its stage-flow representation (see an schematic flowchart in Figure 8.1), the proposed approach follows that of a classical identification system in computer vision. For a previously segregated

object, it consists of: a characterisation stage, which is common for both the training and testing stages; a training stage; and a test stage. However, the operation of each stage presents some peculiarities that—for better understanding—should be stated prior to their detailed description.

Characterization stage

In this study, the characterisation process is region-based and locally-aligned according to singular-points. Descriptions are extracted around object's singular-points which, as generally accepted, are the most repeatable object's evidences in the presence of moderate object rotations and scale changes.

In order to characterise these singular-points, a description support area around them should be used. In existing two-dimensional (SIFT [Lowe, 2004], SURF [Bay et al., 2006] , GLOH [Mikolajczyk et al., 2005], DAISY[Tola et al., 2010]) and three-dimensional (SHOT [Salti et al., 2011] , CSHOT [Salti et al., 2014]) descriptors of singular-points, this area is fixed for every singular-point, both in its shape and in its size.

If the knowledge is acquired from holistic samples—which is the common approach in knowledge modelling—the description support may spread along the whole object, including information that may not be present in occluded instances in the test stage. To solve this problem we propose to restrict the support area to the object's part boundaries.

In order to perform this automatically, we rely on a contour-based RS approach (Arbelaez et al. [2011]) and use, for the description, information of just the pixels (or voxels) belonging to the region containing the singular-point. In Arbelaez et al. [2011], regions are extracted by applying a threshold on the intensity of boundary information of the scene, and such intensity may not be stable among different object's views, mainly in the eventuality of inter-object occlusion. In order to reduce such instability, we opt for a conservative solution: we extract regions at different levels of coarseness—by applying several increasing thresholds on the boundary information— yielding successively smaller regions. This solution is motivated by a trade-off between description capacity and stability. On one side, small regions tend to be less descriptive, as they include less information and these are prone to be more affected by image noise; however, the smaller they are, the higher the probability of finding them in the test stage as non-occluded object evidences. On the other side, attending to the higher characterisation capacity of bigger regions, these are also extracted under the expectation that their boundaries are also partially conserved for different views and under different occlusion situations.

Therefore, each singular-point is described several times, one for each partition at each coarseness level; or, the other way around, a region description may be locally-aligned several times (as descriptions are extracted w.r.t. the local reference frame defined by a singular-point), one per each singular-point in the region.

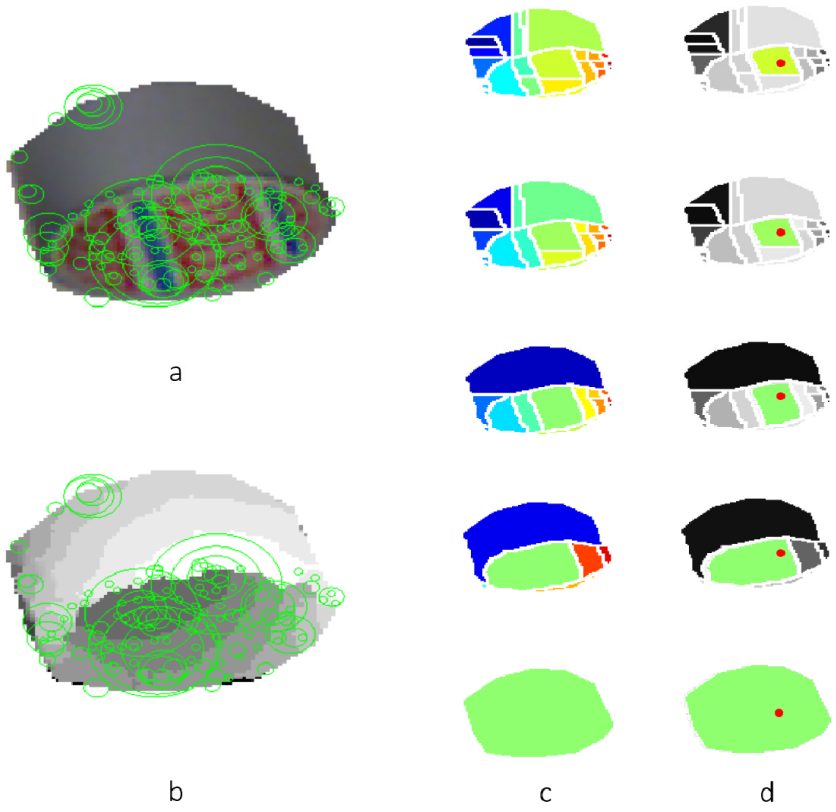


Fig. 8.2. Characterization stage. Singular-points are extracted by means of a scale-space analysis of the chroma and depth information. Points are shown on the colour (a) and depth (b) images of an instance of the *tape* object: the bigger the radio the bigger the scale at which these are detected. The object is then partitioned (c) in several coarseness levels ($E = 5$ in the figure, from top to bottom). A given pixel (the red dot in column d) may belong to a different region at each coarseness level.

This region-point duality can also be understood as a way to overcome a potential lack of representativeness of the training data. One cannot predict which parts of an object will be visible if its occluded—in fact, which parts or regions and which singular-points—. However, occlusions can be bypassed by individually characterising the object parts. Then, if for a new object instance some of the parts are recoverable—those not occluded and aligned with non-occluded singular-points—the whole object may be—at least partially—identified. Figure 8.2 illustrates this concept whereas the process is explained in detail in section 8.4.

Training stage

All of the extracted descriptions for all the training object instances need to be organised and grouped in order to find common descriptions for common object parts. There might be thousands of object descriptions for each object instance, as due to the nature of the characterisation

process, different parts of the same object might be described differently as might be aligned to different singular-points. In order to manage this in an automatic fashion we rely on an unsupervised learning and organising process. This process is able to: arrange highly dimensional data in a manageable structure; store common descriptions together; generate classification boundaries in a (highly) multi-dimensional description space.

From the training process—explained in detail in section 8.5—arises a sheet-like two-dimensional neural map. Each neuron in the map represents—through a description vector of the same dimension as those used for training the map—similarly described object parts—let us call these parts: object *evidences*—. These *evidences* are expected, but not forced, to represent the same object part under the same local-alignment. We call this neural structure: the model or the *SOM* (for the technique used for its training).

The arrangement of neurons in the model is established by taken into account the inter-similarities amongst the neurons. Similar neurons associated to the same modality—i.e. representing similar *evidences*—are placed close in the two-dimensional representation space. The knowledge associated to an specific object can be stored in several non-adjacent neurons in this representation space, each one storing the characterisation of an aligned part of the object. A particularity of the sketched learning process is that, as it is unsupervised, the knowledge arrangement only relies on the description vectors. The overall structure is a three-level knowledge organization: at the first level, singular neurons are expected to convey high similarities (neuron's activation) when compared to test object *evidences* which description vectors are similar to the neuron weight vectors; at the second level, neurons closely located in the two-dimensional representation are prone to be activated (or inhibited) together, then providing robustness to object *evidences* which descriptions slightly differ from the trained descriptions; finally, observing the structure as a whole, an activation/inhibition or excitation pattern (with quantifiable degrees) is obtained. We call this pattern the *signature* of an object *evidence.*

The representation capacity of the model dramatically increases by considering the *signature* instead of just the strongest activated neuron. For instance, a method just considering the strongest activated neuron would be able to represent a number of *evidences* equal to the number of neurons in the model—to say η—. A method considering the responses of all the neurons would be able to represent Q^{η}, with Q representing the number of levels on which the neuron activation intensity is quantified. This is what we call distribute encoding. In our solution we opt for the distributed encoding of the *evidences* but do not define a particular quantification of the activation intensities. Instead, we store the *signature* obtained for each object *evidence* . The *signature* is used for comparison in the testing state. We call the combination of the *signatures*—one per trained object *evidence*—: the *knowledge.*

The training process is described in section 8.5. An sketch of the process is included in the top part of Figure 8.3.

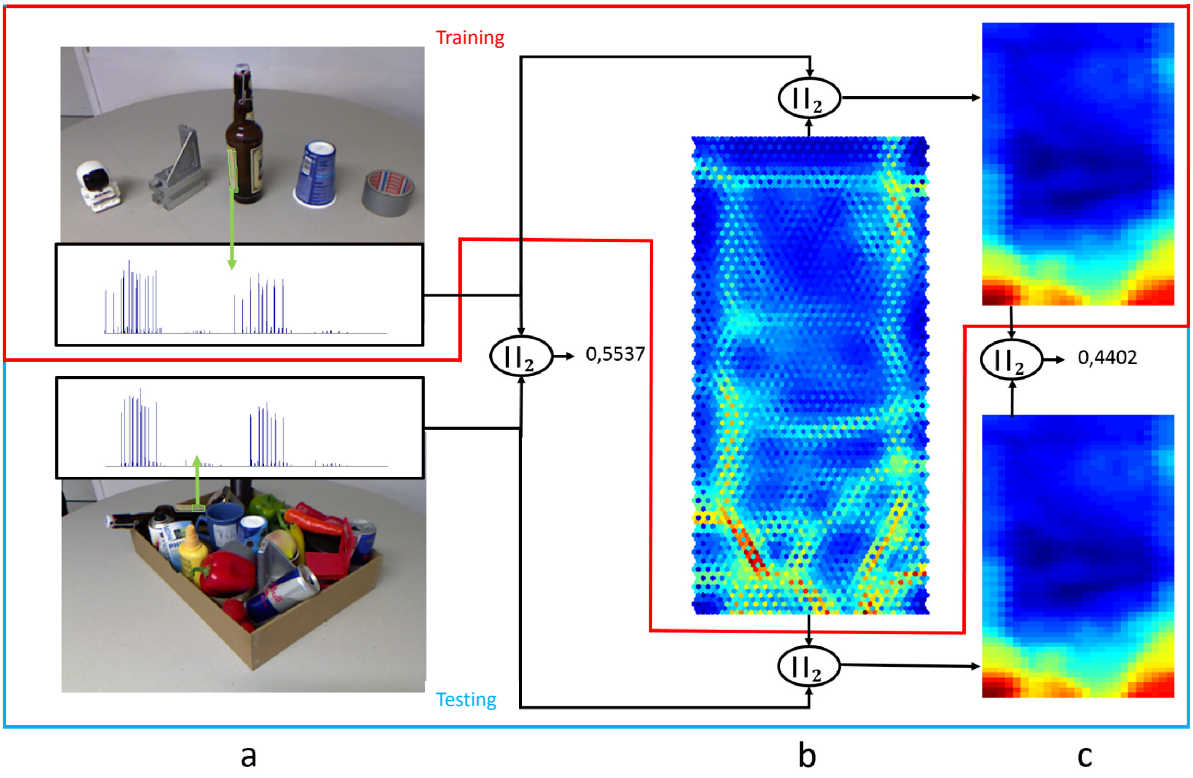


Fig. 8.3. Training / Testing stages. After the characterization stage description vectors are generated (a). Training description vectors from different objects, different object's views and different object partitions and local alignments are used to create a neural model (b). Two excitation patterns or *signatures* are obtained when comparing a training and a testing description vector against each neuron's weight vector in the model (c). The description itself is robust (to some degree) to object rotations due to local-alignment of the descriptions. Note that the similarity between the *signatures* is even higher (of a lower Euclidean distance) than the similarity between the descriptions themselves. Qualitatively, in the example, the excitation patterns are indeed quite similar even when they represent different partitions at different level of coarseness of the *bottle* object.

Testing stage

A straightforward way to measure the likelihood of a test description being a part of a trained object, would consist of comparing it with all the training descriptions (description-to-description). This is the scheme followed by POI matching methods as SHOT ([Salti et al., 2011]). Based on the described neural organisation, a less-resource demanding alternative would be to search for the best neuron-to-description association in a codebook-like scheme (description-to-model). A distribute-encoding alternative would be to evaluate the similarity of two *signatures*: each training *signature* against each test *signature* (*signature*-to-*signature*).

The three alternatives convey a likelihood score at *evidence* level. As our aim is to identify

objects as full entities, an *evidence*-to-object association is required. To this aim, in the proposed approach we follow a labelling procedure: every testing description is scored by *signature*-to-*signature* comparison with the training *evidences.* This process results in a scoring vector which stores the likelihood of such testing description being a representation of an *evidence* of one of the objects. All the testing descriptions that involve an object-point are jointly considered to obtain the likelihood of the object-point being part of a object.

The testing process is described in section 8.6 and is sketched in the bottom of Figure 8.3.

8.4 Feature extraction

Preliminaries

Object-points Let $\mathbf{x}_0$ be a generic object-point—in this chapter both the denomination and the symbol will be used indistinctly for identifying an object-point two-dimensional position, an object-point two-dimensional position with associated depth information and a three-dimensional object-point—. In a Kinect-like scenario, every object-point is defined by its spatial position and its colour and depth information. If not available, the 3D-coordinates of each object-point are estimated by using the internal parameters of the depth sensor—the inverse of its internal calibration matrix—, as described in [Hinterstoisser et al., 2012].

Singular-points Our proposed characterisation technique requires a robust detection of singular-points in the analysed images. For this purpose we use the well established scale-space theory, adapted to the type of images we work with.

Given a RGB colour image registered—in a Kinect-like scenario—with depth information, D, two new image representations are derived: a Lab image, $\mathbf{I}_{Lab}(\mathbf{x})$, which contains the more perceptually uniform $CIELab$ colour representation, and a Dab image, $\mathbf{I}_{Dab}(\mathbf{x})$, built by replacing the luminance information in the Lab image with a normalized version of the depth, D.

A vectorial scale-space representation, $\mathbf{L}_{Dab}(\mathbf{x};t)$, of $\mathbf{I}_{Dab}(\mathbf{x})$ is generated via convolving every image band with a Gaussian kernel, $G(\mathbf{x};t)$ with $\sigma = \sqrt{t}$, so that:

$$\mathbf{L}_{Dab}(\mathbf{x};t) = (L_D(\mathbf{x};t), L_a(\mathbf{x};t), L_b(\mathbf{x};t)) \tag{8.1}$$

, with

$$L_D(\mathbf{x};0) = \mathbf{I}_D(\mathbf{x}); L_D(\mathbf{x};t) = \mathbf{I}_D(\mathbf{x}) * G(\mathbf{x};t); \tag{8.2}$$

, for the D band, and similarly for the a and b bands. The scale dimension is sampled for $t_\iota = \iota\sigma^2, \iota \in \mathbb{Z}$, so that this discrete version of $\mathbf{L}_{Dab}(\mathbf{x};t)$ can in practice be obtained by repeatedly convolving with a small fixed Gaussian kernel, $G(\mathbf{x};\sigma)$.

Singular-points are then declared just for object-points, $\mathbf{x}_\psi$, which are either local maxima or minima (in a spatial circular window of radio $3t_\iota$ per scale ι, and all scales) of the normalized

Laplacian of a band of $\mathbf{L}_{Dab}(\mathbf{x};t)$:

$$\mathbf{x}_{\psi} = argmaxminlocal_{(\mathbf{x})}\left(\nabla^2_{norm} L_D(\mathbf{x};t)\right) \tag{8.3}$$

, for the D band, and similarly for the a and b bands. We also conserve the information about the scale at which each singular-point is detected ι_{ψ}.

An image or matrix, initially set to zero, with the absolute value of the maximum or minimum of every declared singular-point, from now on the *singularity value* ξ_{ψ}, is kept. In case that a same point corresponds to a local maximum or minimum for more than one band, the maximum absolute value is kept, as we aim to detect the most prominent singular-points in every information band. This process can be efficiently performed by using morphological dilation and erosion processes. an example of singular-point extraction is included in Figure 8.4.

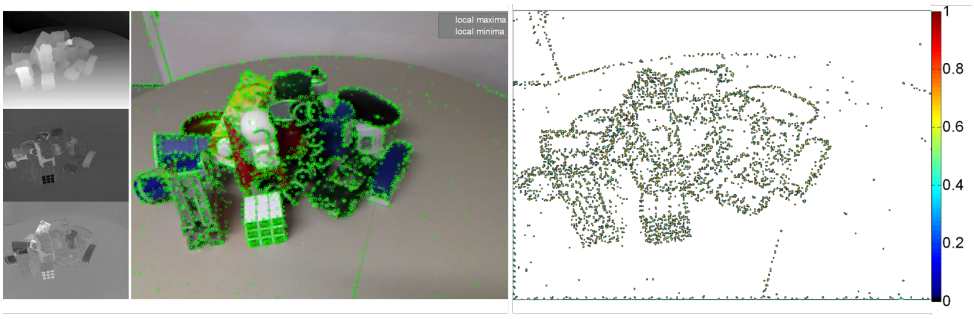


Fig. 8.4. Singular-points extraction from the *Dab* image. Left column: depth information D (top), a (middle) and b (bottom) channels. Middle column: spatial position of detected singular-points. Right column: normalised *singularity value* per singular-point. Singular-points with associated non-scaled value lower than one have been removed for visualization.

Object partition

Objects in the $\mathbf{I}_{Lab}(\mathbf{x})$ image are segmented into regions by means of the approach described in [Arbelaez et al., 2011]. This technique combines colour and texture information to detect the transitions or contours between regions (see chapter 3). One of the main advantages of this segmentation scheme is its hierarchical nature. An object can be divided into one or more regions, according to its internal properties, with each of the regions being defined by its boundary, whereas this boundary is weighted by the strength of the edges that compose it. In other words, the edge strength quantifies the statistical complexity of each segmented region which, in practice, allows to control the coarseness of the partition. By sequentially applying an increasing threshold on the edges strengths, several hierarchical partitions of the object might be obtained. In our experiments, we propose to uniformly sample the edges strengths inside an object in E levels or partitions, forcing the last to be the holistic level, i.e a single region containing the whole object (see top of Figure 8.5). Each region in each partition defines an area—from now on, the *region area*—over the $\mathbf{I}_{Lab}(\mathbf{x})$ image; several singularities might be

detected on the *region area*; whereas every object-point in the *region area* has associated a 3D-coordinate.

Region characterization

We have tested two alternative characterization schemes: one relies on regional descriptions that combine only colour and depth information; and the other describes a region's three-dimensional mesh. Both descriptors are based in successfully State-of-the-Art studies that have proven their effectiveness when used for either the reconstruction or identification of 3D scenarios.

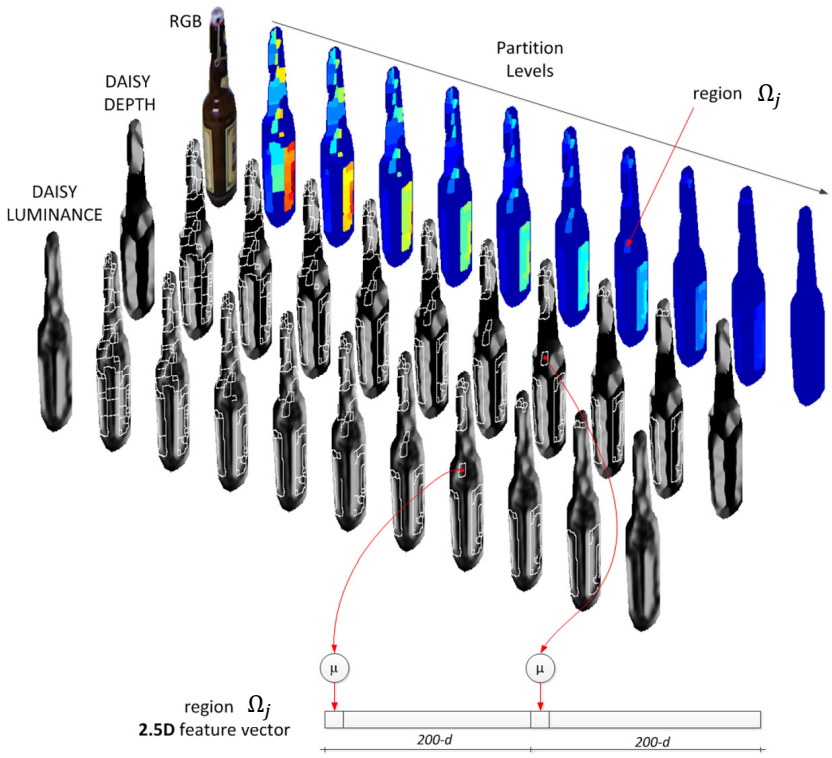


Fig. 8.5. The R-DAISY descriptor. Feature-vector extraction for one instance of the object *bottle*. The original RGB image is converted to its Lab representation and repeatedly partitioned into E coarseness-levels (E=10 in the example). Each region in a partition defines *region area* over the luminance and depth DAISY descriptions—for visualization purposes only the first bin of each DAISY description is shown; depth in the second row and luminance in the third row of bottles—. The feature-vector for a region Ω_j is obtained by concatenating the median value, μ, of the descriptions inside the area defined by the region. Note that, as this description scheme does not rely on the extraction of singularities, it obtains a single description per region.

A luminance and depth regional description. Every region in every object partition is individually characterised by the distribution of the luminance and of the depth information

inside the *region area*. To this aim, we use the DAISY descriptors as proposed in [Tola et al., 2010]. These have proven to be robust to illumination and to moderate perspective and scale changes whereas they can be extracted more efficiently than alternative descriptors.

In our experiments, we respect the standard configuration of DAISY, and then achieve a 200 dimensions (200-d) description vector for the luminance band and another 200-d description vector for the depth band. To provide robustness respect to object rotation, descriptions are extracted w.r.t. the orientation of the maximum axis of the smallest ellipse that circumscribes the region.

In order to give a description at region level, we characterise a region by the concatenation of two vectors, the first one being the median value of the *region area* luminance description and the second one the equivalent for the depth description.

This process leads to a set of 400-d feature-vectors for every object partition, which results in a variable number of feature-vectors per object—one for each region at each partition—.

An example of the object partition and characterisation processes is sketched in Figure 8.5.

This description is extracted once per region, i.e. it is independent of the singular-points, and it is described by a DAISY-like scheme, thus we call it region-masked-DAISY or R-DAISY.

A three-dimensional regional description. In [Salti et al., 2014] authors detail the operation of a 3D descriptor which compiles, via histograms, the information in the neighbourhood—3D support zone—of a 3D singular-point. The histograms are organised in a signature-like structure, by quantifying the spatial position of points in the support zone in a predefined set of 3D volumes. To this aim, the support zone—a sphere of a configurable radio R—is divided into sectors whose ranges are defined in terms of azimuth, elevation and radio. Then, normal and colour vectors of object-points inside each sector contribute to the two histograms describing it and—through a quadrilinear interpolation process—to the neighbouring sectors. This is the so called SHOT (Signature of Histograms of Orientations) descriptor. SHOT is extracted w.r.t. a local reference frame (RF) per described singular-point, then providing invariance to translations and rotations and robustness to noise in the description itself. The local RF is computed from the subset of object-points inside the support of the described singular-point by eigen-value decomposition. Then, these object-points are rotated to the local RF before the geometrical quantization and the histogram-based description. Authors propose to describe a singular-point according to the normal and the texture distribution in its support zone. Half of the SHOT description is devoted to describe the normals by quantifying the cosine of the angle between the normal of the singular-point and the normal of each object-point in the support zone. The other half includes the texture information, which is similarly obtained by quantifying the L1-norm between the *Lab* colour vector of the singular-point and that of each object-point in the support zone . The algorithm is detailed in [Salti et al., 2014].

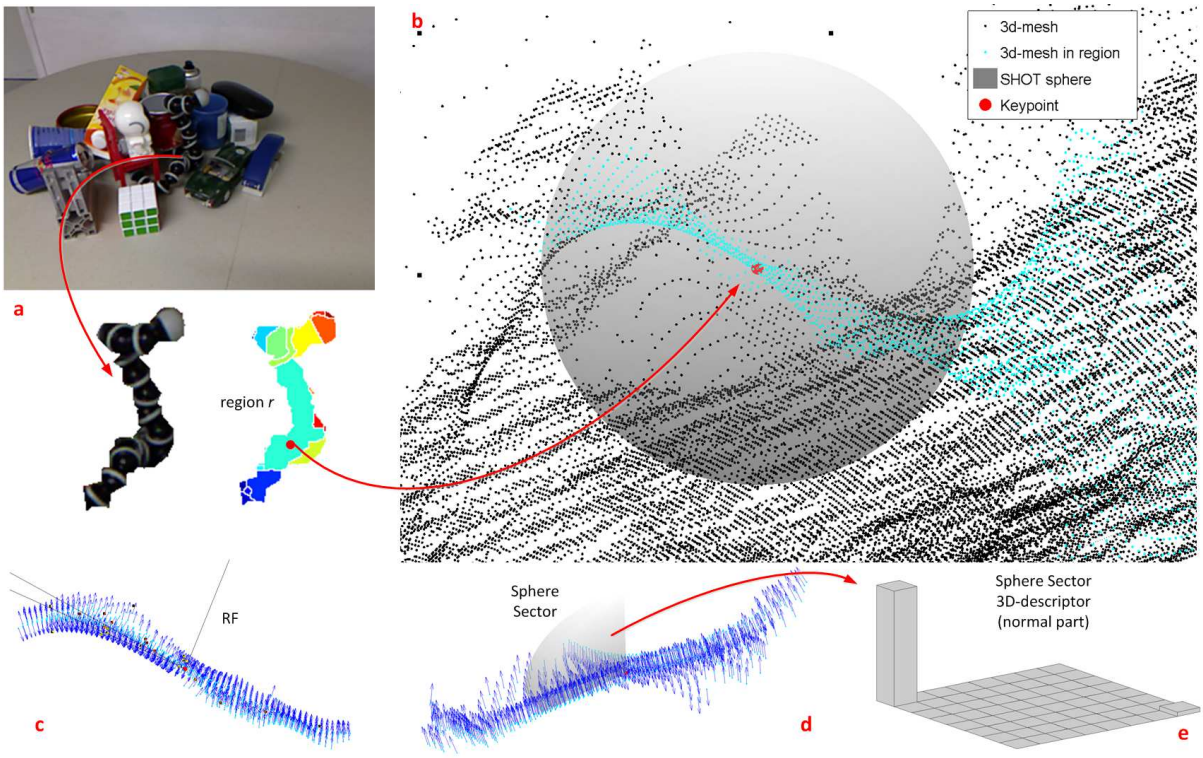


Fig. 8.6. The R-SHOT descriptor. a. The object is partitioned into E coarseness-levels (only the 5^{th} level is shown in the figure). b. For a given singular-point (red point), its associated region in the partition defines a subset of 3D-coordinates (light blue points) over the whole image mesh (black points). Note that this masking strategy is able to avoid the use of points of a different object part in the SHOT description (some of the black points inside the sphere). c. Singular-point's neighbouring object-points (in the image these are plotted with associated normals) are used for the RF estimation. d. The RF is used for the computation of the local coordinates. e. Finally, normals in each sphere sector are described by a two dimensional sample-normalised histogram (in our configuration it is composed of 8x8 bins). Note that, as the singular-point is here an inflection point between surface normals, the description is polarised. Only the extraction of the normal part of the descriptor is illustrated. The process is equivalent for the colour part.

We propose to modify the SHOT descriptor to include region-level information. In our proposal, we also extract SHOT descriptions at the singular-points but the support zone is defined by a double-conditioned process. As in the original SHOT, only object-points inside a sphere of radio R—with R being function of the scale at which the singular-point is detected—are considered; however, only object-points belonging to the same *region area*—hypothesised object part—of the singular-point are used for both the local RF estimation and the description. This process aims to avoid the inclusion in the singular-point description of information that does not correspond to the object part which contains the singular-point. Hence providing robustness to

the description if, in a test instance, the object is partially occluded.

However, this proposed region-based inhibition process comes at a cost: regions in the first or more-detailed levels of the partition are mainly homogeneous in colour and depth (see Figure 8.5). In other words, they might present similar *Lab* and normal vector values in the support zone. Therefore, both the cosine of the normal angle and the L1-norm of the *Lab* vectors would be prone to return similar values. In practice, this would result in a histogram description where only a few bins would be filled, independently of the normals and colours in the support zone. Consequently, SHOT descriptions would just differ in the geometrical points distribution. In our initial experiments we concluded that this information was insufficient to identify the objects.

In order to overcome this problem we propose to modify the information contained in the SHOT histograms. The uni-dimensional histogram that accounted for the normals cosine was replaced with a two-dimensional histogram that quantified the elevation and azimuth angle of the normals in the support zone Similarly, the histogram quantifying the L1-norm was replaced by another two-dimensional histogram, now describing the azimuth and elevation angles of the *Lab* vectors. This process is sketched in Figure 8.6.

We keep the standard geometrical configuration of SHOT, then dividing the sphere in 32 sectors with the following configuration: 8 partitions in azimuth, 2 partitions in elevation and 2 more partitions for the radio. In our proposal, the azimuth component of normal and colour vectors is quantified in 8 bins, whereas 8 more bins are used for the elevation component, which overall results in a 2048-d feature-vector for normals description and another 2048-d vector for colour description.

The so-built descriptor uses region information to describe each singular-point through a SHOT-like scheme, thus we call it region-masked-SHOT or R-SHOT.

8.5 Organising the objects knowledge

The self-organised neural structure

We propose to organise the training descriptions obtained for one of the characterization methods described in section 8.4 in two neuron-based structures (one per method) that will constitute the model part of the *knowledge*. To this aim, we have selected the self organising map (SOM) [Kohonen, 1990] as an appropriate framework for training.

Description and biological background. A SOM is a discrete two-dimensional representation or *map* of the training data; each cell, containing a *weight* or feature vector with the same dimension as the input description vectors, represents a neuron. SOMs are trained iteratively under an unsupervised competitive learning: a single winning neuron is determined at each iteration. At the end, the weights of every neuron are adapted so that every neuron will *respond*

more strongly to input-like descriptions.
The fundamental of the SOM is the soft competition between the neurons; not only one neuron (the winner of every competitive learning process) but also its neighbours are updated in every iteration. The SOM training derives in a fully connected single-layer linear network, where the resulting structure—the output layer or the *map*—is organized in a two-dimensional sheet-like arrangement of nodes or neurons, where these are *functionally* connected to their neighbouring neurons. This idea agrees with the modelling of basic information processes in the cortex, where synaptic connections among the neurons are built depending on the type and frequency of sensory stimuli. Through such process neuron-groups become sensitive to specific patterns encoded in perceived signals, and neighbouring neurons tend to learn similar patterns. This makes the SOM a useful tool for visualizing and storing the hypothetical engram, or neuron connections, described by the Hebbian modelling.

Input data. Let N be the number of objects classes to identify; let O_i be the set of training instances for object i, which overall results in N training sets; and let m_j, $j = 1...card(O_i)$ be the number of n-dimensional description vectors characterising each object—one per region for the R-DAISY descriptor or one per region singular-point combination for the R-SHOT description—. This adds up to $M_i = \sum m_j$ description vectors $\mathbf{f}_{i,k} = \{f_{i,k,1}, f_{i,k,2}, ..., f_{i,k,n}\}$, $1 \leq k \leq M_i$, $1 \leq i \leq N$ for each set, and up to $M = \sum M_i$ description vectors for training. Description vectors are organised in an Mxn matrix. This matrix, after variance-based column-wise normalisation, is the training data to generate each SOM.

Determining the map shape. Several parameters should be specified before the training: the number of neurons, the dimensions of the map, the map lattice and the geometry to fill the map. In order to keep a good balance between the flexibility of the resulting map and the computational simplicity of the training stage we follow the advice in [Vesanto, 2005] and define the number of neurons (η) as a function of the number of input descriptions. These neurons are to be organised in a rectangular lattice; its sides (HxW) are determined, again following [Vesanto, 2005] according to the the ratio between the two bigger eigenvalues of the training data. Finally, we follow the default geometry in [Vesanto, 2005] and fill the lattice with hexagonal-shaped neurons.

Training algorithm. The SOM training process is very similar to the one performed by the C-means algorithm (see chapter 3)—with the neighbouring-learning exception—. As neurons weights are of the same size of the description vectors , the weight vector of neuron γ^{th} can therefore be denoted: $\mathbf{w}_\gamma = \{w_{\gamma,1}, w_{\gamma,2}, ..., w_{\gamma,n}\}$. Weight vectors for every neuron $1 \leq \gamma \leq \eta$ in the map, located in the (x_γ, y_γ) position of the rectangular grid, are first initialised by linear spanning of the two eigenvectors of the training data associated to their two bigger eigenvalues.

For each input feature vector $\mathbf{f}_{i,k}$ its best matching (winning) neuron $\dot{\gamma}$ in the map is found by comparing, via the *L-2* norm, the feature vector with the neurons weights:

$$\dot{\gamma} = argmin_{\gamma} \left\{ \|\mathbf{f}_{i,k} - \mathbf{w}_{\gamma}\|_2 \right\} \tag{8.4}$$

, where w_{γ} is the weight associated to neuron γ.

Weights of the wining neuron and of its neighbouring neurons are then updated at each learning iteration $\kappa + 1$ by:

$$\mathbf{w}_{\gamma}(\kappa + 1) = \mathbf{w}_{\gamma}(\kappa) + h_{\dot{\gamma}}(\kappa) \|\mathbf{f}_{i,k} - \mathbf{w}_{\gamma}\|_2 \tag{8.5}$$

, with $h_{\dot{\gamma}}(\kappa)$ being a Gaussian kernel profile centred at the spatial coordinates of the wining neuron $\dot{\gamma}$: $(x_{\dot{\gamma}}, y_{\dot{\gamma}})$ in the rectangular grid, such that:

$$h_{\dot{\gamma}}(\kappa) = \beta(\kappa) exp(- \|(x_{\dot{\gamma}}, y_{\dot{\gamma}}) - (x_{\gamma}, y_{\gamma})\|_2^2 / 2\sigma^2(\kappa)) \tag{8.6}$$

, where the learning rate $\beta(\kappa)$ and the kernel bandwidth: $\sigma(\kappa)$ are decreasing functions of learning time with $0 \leq \beta(\kappa) \leqq 1$.

Labelling the SOM. When the learning process converges (say at κ_{end}) every neuron in the map has an associated weight vector: $\mathbf{w}_{\gamma}(\kappa_{end})$ which is tuned to the training data. In order to use the map as codebook, each neuron should be tagged as a representative of at least one of the object classes. The simple labelling approach that we use consist of:

i) For each labelled training description vector, we extract its best matching (winning) neuron in the learned SOM:

$$\dot{\gamma} = argmin_{\gamma} \left\{ \|\mathbf{f}_{i,k} - \mathbf{w}_{\gamma}\| \right\} \tag{8.7}$$

ii) This neuron inherits the description vector's label, i.e. the label of the training object instance from which it is was extracted, so that a neuron can be labelled with several labels (it is the winning neuron for description vectors with different labels) as well as several times per label (it is the winning neuron for several description vectors from the same object class).

iii) If a unique label per neuron is required (for the codebook-like scheme), a neuron is labelled with its most repeated (frequent) label.

The so-obtained SOM constitutes the model in the proposed scheme. Two models have been trained up to this point. A SOM trained with the R-DAISY descriptions ($SOM_{R-DAISY}$) and a SOM trained with the R-SHOT descriptions (SOM_{R-SHOT}).

Activation / inhibition responses: *signatures*. Every object to identify is characterized during the training stage by a set of *signatures*,

$$S_i = \{s_{i,1}(x,y), ...s_{i,k}(x,y)...s_{i,M_i}(x,y)\} \tag{8.8}$$

A *signature* $s_{i,k}(x,y)$ for object i is defined as the response of a description vector $\mathbf{f}_{i,k}$ in O_i to the corresponding model. This process leads to M_i *signatures* (one per training *evidence*).

A *signature* is computed as the *L-2* norm of a description vector to the weight vector of every neuron in the model, which ends up in a HxW scalar matrix:

$$s_{i,k}(x_\gamma, y_\gamma) = \|\mathbf{f}_{i,k} - \mathbf{w}_\gamma\| \tag{8.9}$$

Every *signature* is labelled with the object-class of the description vector (e.g. class i for $\mathbf{f}_{i,k}$, $1 \leq k \leq M_i$) used to generate it and with the segmentation level on which the description vector has been extracted.

Figure 8.8 includes a couple of *signatures* for visualisation (see section 8.7 for details). Dark areas, local minima or *holes* in a *signature* indicate neurons tuned (high similarity, low distance) to the description vector, whereas bright areas, local maxima or *peaks* indicate neuronal inhibition.

The whole process results in a $N-$length superset of *signatures*: $\mathbf{K} = \{S_1, ..., S_i, ..., S_N\}$ which, together with the model (the SOM), constitutes the learned *knowledge* for a given description method.

The described scheme: segmentation + characterization + model construction, allows to describe every training description vector with an excitation pattern to the common model. Note that these responses, the *signatures*, are a sort of templates that integrate the activation and inhibition *intensities* of all the neurons in the model to description vectors of the trained objects.

8.6 Identifying object instances

In a test scenario, given an object instance o_{test}, our approach consists of assigning to each object-point (a pixel or a voxel) a value indicating its score or likelihood of belonging to each of the training objects. The aim of this section is to describe how these scores are obtained.

The identification process is divided into three stages:

1. A test object instance is characterised, leading to a set of descriptions vectors $\{\mathbf{f}_{test}\}$ (one per *evidence*) which are compared against the SOM, obtaining a set of *signature* $\{s_{test}(x,y)\}$.

2. These *signatures* are then compared against the trained *signatures* and a final set of scores *evidences*-to-object are obtained $\{p(test,i)\}$ for each evidence and each trained object i.

3. The so-obtained scores are combined to derive a result at object-point level. Note, that the number of scores (one per *evidence*) available to identify an object instance varies with the selected description method (see Figure 8.7).

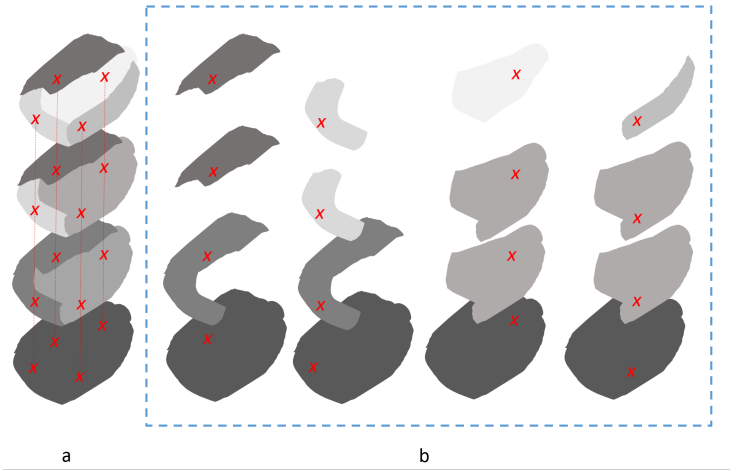


Fig. 8.7. Number of *evidences* per object instance. Synthetic example. (a) An object instance is successively partitioned into $E = 4$ coarseness levels. Same-shaped regions at different partition levels are here considered different regions; hence there are a total of 10 regions in the hierarchical partition. These are here represented by different grey levels in each coarseness level. In parallel, four singularities are extracted on the object instance. Represented by red crosses in the Figure. A total of $m_j = 10$ *evidences* are described following the R-DAISY description scheme (one per region) and a total of $m_j = 16$ *evidences* are described following the R-SHOT description scheme , one per region-singularity pair as showed in (b).

Extraction of the set of description vectors

The procedure to obtain the list of scores for every object-point can be sketched as follows, depending on the two considered description schemes.

R-SHOT description. The object instance is hierarchically partitioned into E levels as described in Section 8.4, and singular-points are detected according to the technique described in Section 8.4. For each detected singular-point at every region, a test *signature* $s_{test}(x, y)$, or response of the SOM to the corresponding description vector $\mathbf{f}_{test}$, is computed; this test *signature* is compared with the set of trained *signature*, obtaining a set of similarity scores; the maximum score for the subset of an object's training *signature* defines that object's score for the test *signature*; this results in a set of scores or likelihoods of a description being extracted from every trained object. In order to obtain these likelihoods for all the object-points (voxels) in the object instance, these scores are first propagated to the regions obtained through the hierarchical segmentation process.

R-DAISY description. After the hierarchically partition—which is common for both description schemes—a description vector is directly extracted for every region in every partition. The scores of their corresponding *signatures* respect to every trained object do already provide region-level information. Finally, every object-point (pixel), which might belong to up to E different regions, is assigned a likelihood of belonging to each trained object by combining the likelihood obtained for each of these regions. Let us formalise these stages.

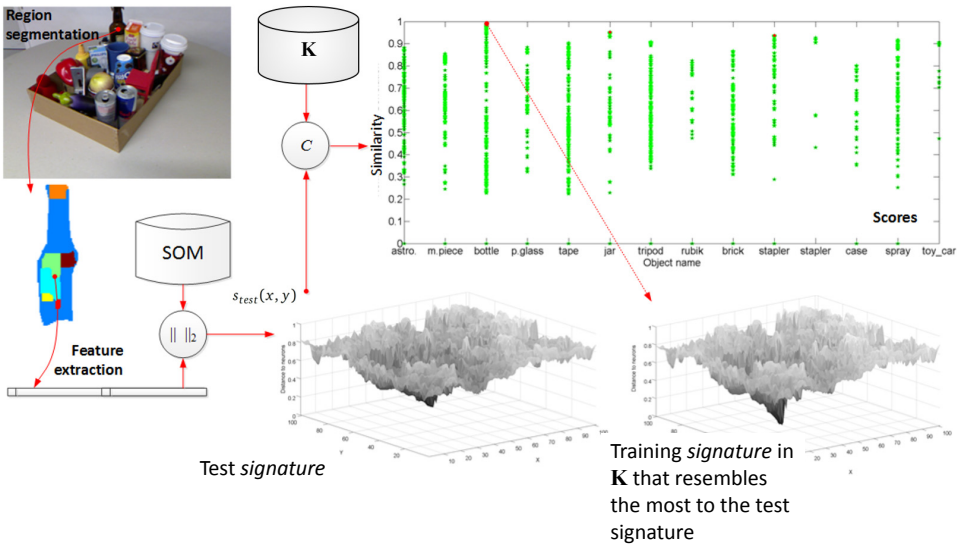


Fig. 8.8. Comparison of a test object *evidence* described with the R-DAISY descriptor against the *knowledge*. A new object instance is segmented into E coarseness-levels (only the 5^{th} level is shown in the figure). An object *evidence*, here represented as an object region, is described by the R-DAISY descriptor as explained in section 8.4. The so-built description vector is compared to every neuron in the SOM, leading to a *signature*: $s_{test}(x,y)$ for each region. This *signature* is then *compared* (C) to every stored one in the knowledge **K**. Ideally, the *signature* would result in neuronal responses similar to those obtained during its object's training stage.

Comparing evidences to the model: scoring each object *evidence*

Let $s_{test}(x,y)$ be the *signature* obtained for an *evidence* description vector $\mathbf{f}_{test}$ extracted in the test stage. In order to evaluate the similarity between this *signature* and every training *signature* in the **K** superset, any *template-matching* technique would convey adequate results (normalized cross-correlation, sum of absolute differences, covariance distances, etc.). However, we instead aim to simplify, and hence, improve the efficiency of the comparison. To this aim, we rely on a likelihood measure which decreases exponentially with the L2-norm between the test and the training *signatures*:

$$p(s_{test}, s_{i,j}) = exp(-\|s_{test} - s_{i,j}\|_2) \tag{8.10}$$

Under the same simplification premise, the set of scores assigned to $s_{test}(x, y)$, one respect to every object i:

$$\{p(s_{test}, i), i = \{1, ..., N\}\} \tag{8.11}$$

, which can be understood as the likelihood of the *signature* corresponding to that object, results:

$$p(s_{test}, i) = max\left[p(s_{test}, s_{i,j})\right], j = \{1, ..., M_i\} \tag{8.12}$$

, i.e. the likelihood of $s_{test}(x, y)$ being the response of the SOM to a description vector extracted from an instance of object i is the maximum likelihood obtained when comparing the test *signature* with all the training *signatures* tagged as object i.

This process is coherent with the bottom-up flow of information described in section 8.2 and is the key principle to enhance system's robustness to occlusions. Different parts of an object may be described with completely different descriptions if these are differently aligned. These parts constitute different *evidence*s. We search for *evidence*-to-*evidence* associations.

Exporting singularity likelihoods to regions

In the R-SHOT description scheme, a test region generally includes a variable number Ψ of singular-points $\mathbf{x}_\psi, \psi = \{1, ..., \Psi\}$; hence, the region is described with Ψ description vectors $\mathbf{f}_{test,\psi}, \psi = \{1, ..., \Psi\}$, i.e., characterized with Ψ test *signatures* $s_{test,\psi}, \psi = \{1, ..., \Psi\}$.

An strategy to combine the scores of each contributing *signature* is required. A straight procedure would be to follow again a non-maximum-suppression process, i.e, keep the highest likelihood per region independently of the singularities to which the description vectors are aligned. However, we consider more robust to include the singularity value $\xi_\psi, \psi = \{1, ..., \Psi\}$ of each singular-point in the process.

Let us define Ω_e as a region obtained at the e coarseness-level partition ($e = \{1, ..., E\}$) and let $\{p(s_{test,\psi}, i), \psi = \{1, ..., \Psi\}\}$ be the set of likelihoods of the test *signature* corresponding to object i obtained for each of the Ψ singular points that lie on Ω_e. Then the likelihood of Ω_e being a region of object i is computed as:

$$p(\Omega_e, i) = \sum_\psi \dot{\xi}_\psi p(s_{test,,\psi}, i) \tag{8.13}$$

, for the R-SHOT descriptor, where:

$$\dot{\xi}_\psi = \frac{\xi_\psi}{\sum_\psi \xi_\psi} \tag{8.14}$$

, are the normalised singularity values.

This scheme aims to aggregate the scores obtained for every alignment of the region. The score is biased towards the singular-points with highest relative singularity value, as these are supposed to represent more stable cues under point-of-view variations.

The set of scores region-to-object can be expressed as:

$$\{p(\Omega_e, i), i = \{1, ..., N\}\} \tag{8.15}$$

In the R-DAISY description scheme, a test region is described by a single description vector, i.e. it is characterized by a single *signature*; hence, the set of scores for that region equals the set obtained for its *signature*:

$$\{p(\Omega_e, i), i = \{1, ..., N\}\} \equiv \{p(s_{test}, i), i = \{1, ..., N\}\} \tag{8.16}$$

Exporting region likelihoods to object points

We here defined our approach to export region-level scores to object-points.

Let $e = \{1, ...E\}$ be the coarseness-level of the hierarchical partition process. As aforementioned, a specific object-point $\mathbf{x}_0$ may belong to E different regions, one per coarseness-level: $\{\Omega_e, e = \{1, ..., E\}\}$.

Let us define $p(\mathbf{x}_0, i)$ as the likelihood of the object-point $\mathbf{x}_0$ belonging to an object i; then the set of object-point scores stands:

$$\{p(\mathbf{x}_0, i), i = \{1, ..., N\}\} \tag{8.17}$$

, which are obtained by inheriting the maximum score obtained for $\mathbf{x}_0$ in all the coarseness levels:

$$p(\mathbf{x}_0, i) = max_e \left[p(\Omega_e, i)\right], e = \{1, ..., E\} \tag{8.18}$$

, i.e. inheriting the score from the region that better describes it according to the training data.

Finally $\mathbf{x}_0$ is labelled $\mathcal{L}(\mathbf{x}_0)$ with the object class that maximises the score, i.e., with the object label which to-knowledge comparison produces the highest score:

$$\mathcal{L}(\mathbf{x}_0) = argmax_i(p(\mathbf{x}_0, i)) \tag{8.19}$$

, with associated final score:

$$\mp(\overline{x}_0) = max_i(p(\mathbf{x}_0, i)) \tag{8.20}$$

This process applies, with different inputs, for both the R-DAISY and the R-SHOT descriptors.

8.7 Case of example: Severe-occluded objects identification.

Dataset description

In order to evaluate the discrimination capability of the selected region-driven descriptors (R-DAISY and R-SHOT) and the ability of a SOM-derived set of *signatures* to represent an object's knowledge, we propose to evaluate our proposed technique with the challenging data-set presented in [Potapova et al., 2011].

The training part of the data-set is composed of 50 instances of 14 full-represented objects distributed in 10 images (see Figure 8.9). In practice, this entails that there are only 2-to-8 short-varied samples per object for the training stage.

The testing part of the data-set includes a total of 448 object instances of the same 14 objects distributed in 75 images, which also include many other unlabelled—and hence untrained—objects which are not analysed in the current version of the proposed method. These 14 objects are quite varied in appearance, going from highly textured—*glass*—, through middle textured—*astronaut, carton of juice, cup, toy car, rubik cube*—to untextured or flat objects—*case, stapler*—, these also including metallic objects—*metal piece, spray*—. The dataset also comprises a highly deformable object—*tripod*—, two crystal objects—*bottle*, *jar*—and a non-compact object—*tape*—.

Depth information—due to the Kinect capture technology—is missing in some areas at the crystal: *bottle*, *jar* and the metallic objects: *metal piece*, *spray*. In our experiments we have used an interpolation method relying on a 8-connected spring metaphor to estimate missing depth information.

This data-set is, to our knowledge, first used for the evaluation of an object identification system—in [Potapova et al., 2011] was used instead for detection of saliency cues—. In our opinion, this is due to a couple of challenging factors that disregards its use:

1. As aforementioned, the training data-set only contains 10 frames where the objects appear isolated—see first row on the left part of Figure 8.9—.

2. In the test data-set the trained objects are severely occluded then leading to unpredictable appearances of these objects—some examples are shown in the second row of Figure 8.9—.

This data-set perfectly adapts to the target scenario that we consider, as the training data-set is small, occlusions are varied and natural—not synthetically generated—and object's spatial location is annotated, which allows to bypass the segregation stage—alternatively, the solutions proposed in [Lyubova and Filliat, 2012] or [Gallego and Pardàs, 2014] can be used—. This condition permits the evaluation of the identification method independently of this stage.

Alternative state-of-the-art data-sets, as those used in [Salti et al., 2014] and [Hinterstoisser et al., 2012], do not contain severe occluded objects; then, they are not suitable for evaluating

the hypothetical advantages of the proposed approach.

Experiments description

We propose to compare the behaviour of three descriptors: the State-of-the-art shape and colour version of the standard **SHOT** descriptor—sometimes named CSHOT—configured as proposed in [Salti et al., 2014], the proposed **R-DAISY** descriptor and the proposed **R-SHOT** descriptor. Remember that there is one **R-DAISY** descriptor per region and might be several **R-SHOT** descriptors per region (one per singular-point).

The efficient template-based approach of [Hinterstoisser et al., 2012] has been discarded for comparison as authors declare it to be only robust to less than 20% occlusion situations, and uses CAD models—unavailable in this case—in both the training and testing stages.

We have carried out identification experiments under three configurations.

1. Configuration c.1: neither the model (SOM) nor *signatures* are used for identification, i.e. only descriptions are used for training and testing (descriptor-to-descriptor).

2. Configuration c.2: the model (SOM) is used but *signatures* are not used. Testing is performed comparing description vectors with the SOM neurons, following a codebook-like scheme (description-to-neuron).

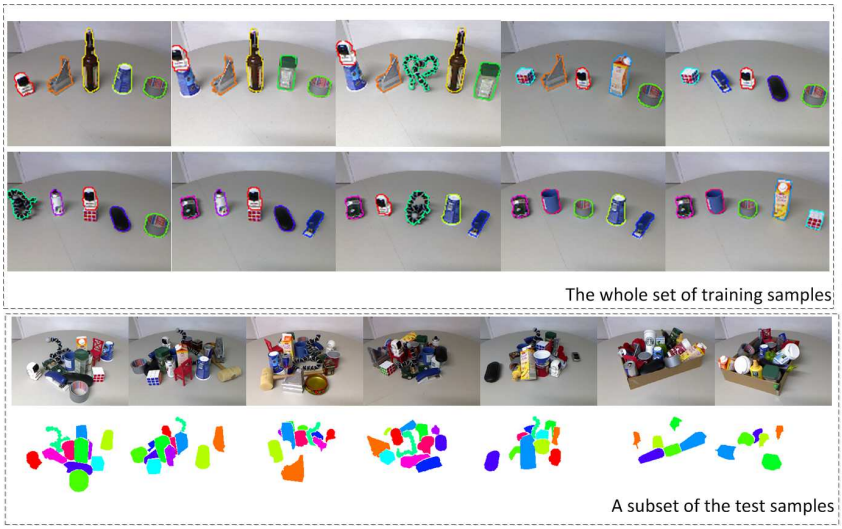
3. Configuration c.3: The model (SOM) and the *signatures* are used for training and testing, which is our proposal (*signature*-to-*signature*).

The aim of this incremental evaluation procedure is to measure the advantages of each of our proposed contributions.

The c.1 configuration aims to compare the identification capability of the proposed descriptors, **R-DAISY** and **R-SHOT**, against **SHOT**; the c.2 one is devoted to measure the contribution of the *knowledge* grouping process performed by the SOM; and the c.3 one aims to evaluate the hypothetical benefits of the whole approach.

Specially, this last configuration aims to evaluate the feasibility of using distributed encoding for the identification of severely occluded objects trained with a small amount of instances. The maximum training object instances is 9 for the *astronaut* object.

The **training stage** for the three descriptors is equivalent; first, each object training instance is characterised by means of each one of the three descriptors; then, a SOM is trained for each descriptor, thus, three different SOMs ($SOM_{R-DAISY}$, SOM_{R-SHOT}, SOM_{SHOT}) are trained and labelled according to section 8.5; finally, by comparing the descriptors with the SOM through equation 8.9 three (one per descriptor type) sets of signs (equation 8.8) are obtained for each trained object.



Object name	Color code	# instances @ training	# instances @ test
Astronaut		8	26
Metal piece		4	43
Bottle		3	16
Glass		4	52
Tape		7	36
Jar		2	37
Tripod		3	28
Rubik cube		3	22
Carton of Juice		2	51
Stapler		4	18
Case		3	34
Spray		2	34
Toy car		3	15
Cup		2	36

Fig. 8.9. The dataset analysed—presented in Potapova et al. [2011]—. Top-left: the whole training set of samples. Note that a very few number of samples are available for training. Additionally, in some cases the samples are captured from the same point-of-view. Right: the objects to identify, a colour code to represent them and the number of training and test instances per object. Bottom-left: Some frames from the test dataset and the associated ground-truth. Observe that objects to identify are severely occluded and even placed inside a box, partially occluding their original contours an inner-structure.

The **testing stage** for the three descriptors is similar; first, each object test instance is described by means of each one of the three descriptors (these descriptions are used in the c.1); then, each descriptor is compared to their corresponding SOM (e.g. a **R-DAISY** descriptor is compared against the $SOM_{R-DAISY}$), leading to a winning neuron (descriptions and SOM are used in the c.2 configuration) and to a *signature* (used in the c.3 configuration).

The **performance** of each configuration (c.1 to c.3) is measured via two analyses. First *Recall*, *Precision* and $F_1 - score$ statistics evaluating object-point classification are obtained for each object at an optimal operation point—which might be different for each descriptor—, hence comparing best-balanced operations. Second, in order to better assess the discrimination capability of the evaluated configurations a confusion matrix is also included.

Results are given for an object identification task, i.e. its operation is only evaluated in test instances of trained objects, thus neither untrained objects nor background samples are included in these statistics.

Computing identification scores

There is one **SHOT** descriptor per singular-point, one **R-DAISY** descriptor per region and one **R-SHOT** descriptor per region-singular-point couple. In order to obtain identification scores for each segmented object, the processes to export scores from singular-points to regions (just for

the **R-SHOT** case) and from regions to object have been already described for the **R-DAISY** and the **R-SHOT** descriptors (section 8.6).

In order to compare the **R-DAISY** and the **R-SHOT** descriptors with the **SHOT** descriptor, as this last just provides scores for singular points, a score exporting process should be defined.

In the test stage Ψ' singular-points: $\mathbf{x}_{\psi'}, \psi' = \{1, ..., \Psi'\}$ are detected on an object instance o_{test}. For the **SHOT** descriptor each of these singular-points is described by a **SHOT** feature vector $\mathbf{f}_{test,\psi'}$ which leads to a *signature* $s_{test,\psi'}$ when compared to SOM_{SHOT}.

Let $\{\xi_{\psi'}, \psi' = \{1, ..., \Psi'\}\}$ be the set of singularity values of these singular-points. Then the likelihood of o_{test} being an instance of object i is here computed as:

$$p(o_{test}, i) = \sum_{\psi} \dot{\xi}_{\psi'} p(s_{test,\psi'}, i) \tag{8.21}$$

, for the **SHOT** descriptor, where:

$$\dot{\xi}_{\psi'} = \frac{\xi_{\psi'}}{\sum_{\psi'} \xi_{\psi'}} \tag{8.22}$$

, are the normalised singularity values.

By means of this process, every object-point $\mathbf{x}_0 \in o_{test}$ in the object instance is equally scored, i.e. for the **SHOT** descriptor results are given at object level (not at region nor *evidence* level). This has a relevant effect in the results for this descriptor. We discuss about this effect in the results discussion subsection.

Finally, object-points are labelled as being part of the object that maximise the score:

$$\mathcal{L}(\mathbf{x}_0) = argmax_i(p(o_{test}, i)), \forall \mathbf{x}_0 \in o_{test} \tag{8.23}$$

, with associated final score:

$$\mp(\mathbf{x}_0) = max_i(p(o_{test}, i)) \tag{8.24}$$

In our experiments, we have observed that this scheme substantially improves the performance of the **SHOT** descriptor , compared to a direct classification of the object. In particular, we have explored and discarded—for their lower performance—two alternative schemes: (1) consider only the probability associated with the most-prominent singular-point; and (2) identify according to the most probable singular-point to object association.

Evaluated configurations

The three configurations propose different strategies for the calculation of the sets of scores per object-point $\{p(\mathbf{x}_0, i), i = \{1, ..., N\}\}$. We here describe the details for each of these strategies. It

is important to remark that the outputs of the experiments carried out are also dependent of the type of description used: as three different descriptions are proposed and these are configured in three different ways; there would be a total of nine different outputs for comparison.

The objective is to obtain the labels $\mathcal{L}(\mathbf{x}_0)$ and associated final scores $\mp(\mathbf{x}_0)$ for every pair descriptor-configuration, in every object-point $\mathbf{x}_0$ in the test stage.

Configuration c.1. In the first configuration, descriptions are matched among themselves—as in a singular-point matching approach—, i.e., the SOM is not used and then no signs are generated. Given a test description vector $\mathbf{f}_{test}$ and the set of description vectors used for the training of the object i:$\{\mathbf{f}_{i,k},\ 1 \leq k \leq M_i\}$, the likelihood of $\mathbf{f}_{test}$ being generated from an instance of object i is computed as:

$$p(\mathbf{f}_{test}, i) = max_k(p(\mathbf{f}_{test}, \mathbf{f}_{i,k})) \tag{8.25}$$

, where:

$$p(\mathbf{f}_{test}, \mathbf{f}_{i,k}) = exp(-\left\|\mathbf{f}_{test} - \mathbf{f}_{i,k}\right\|) \tag{8.26}$$

In order to obtain results per object-point and to label the object-points, an exportation process of so-computed description vector likelihoods is differently performed for each descriptor type.

For the **R-DAISY** descriptor so-computed likelihoods are already at region level, as $\mathbf{f}_{test}$ describes a whole region; thereon, these are exported to object-points by simply using them instead of the *signatures* likelihoods in the process described in section 8.6. For a formal derivation, replace $p(s_{test}, i)$ with $p(\mathbf{f}_{test}, i)$ in equation 8.16 and then apply equations 8.17-to-8.20.

For the **R-SHOT** descriptor each $\mathbf{f}_{test}$ on a region is associated to a singular-point: $\mathbf{f}_{test,\psi}, \psi = \{1, ..., \Psi\}$; so, to convert these to region-scores,description likelihoods are used instead of *signatures* likelihoods in the process described in section 8.6. Afterwards, the process described in section 8.6 is used to derive object-point scores. For a formal derivation, replace $p(s_{test,,\psi}, i)$ with $p(\mathbf{f}_{test,\psi}, i)$ in equation 8.13 and then apply equations 8.17-to-8.20.

Finally, for the **SHOT** descriptor, each $\mathbf{f}_{test}$ on an object instance is associated to a singular-point: $\mathbf{f}_{test,\psi'}, \psi' = \{1, ..., \Psi'\}$. These are exported to object-points by using them instead of the *signatures* likelihoods in the process described in section 8.7. For a formal derivation, replace $p(s_{test,,\psi'}, i)$ with $p(\mathbf{f}_{test,\psi'}, i)$ in equation 8.21 and then apply equations 8.22 and 8.23.

Configuration c.2. In the second configuration, the SOM is used as codebook; then descriptions similarity to the SOM is first evaluated. Given a test description vector $\mathbf{f}_{test}$, the winning neuron for $\mathbf{f}_{test}$ in the SOM is determined by:

$$\dot{\gamma} = argmin_\gamma \left\{\left\|\mathbf{f}_{test} - \overrightarrow{w_\gamma}(t_{end})\right\|\right\} \tag{8.27}$$

As the winning neuron is already labelled with a training object label, say $\mathcal{L}(\dot{\gamma}) = i$, each description vector is associated to just a single object, with score:

$$p(\mathbf{f_{test}}, \mathbf{i}) = \mathbf{exp}(-\left\|\mathbf{f_{test}} - \overrightarrow{w_{\dot{\gamma}}}(\mathbf{t_{end}})\right\|) \tag{8.28}$$

, whereas $p(\mathbf{f}_{test}, i') = 0 \, \forall i' \neq i$.

The process to export scores to object-points and to label them for each descriptor type is then equivalent to the one applied for configuration c.1, but by using this description vector likelihoods instead.

Configuration c.3. The third configuration relies on the use of the SOM responses to the description vectors, i.e. on the *signature*s. This is the approach that has been fully described in this chapter.

System set-up

- **Singularity extraction.** We use $1 \leqq \iota \leqq 5$ scale samples with a Gaussian kernel standard deviation $\sigma = 1$. For the standard **SHOT** descriptor, as [Salti et al., 2014] did not specify the amount of singular-points extracted, we split the total set of singular-points extracted into two subsets according to the absolute singularity values and keep those with the highest values. This process results in an average number of 833.56 singular-points per object instance in the training stage. For the **R-SHOT** descriptors we use the same singular-points extracted for the **SHOT** approach, so that the comparison is unaffected by the singular-point detection process.

- **Object partition.** The edges strengths inside an object obtained through [Arbelaez et al., 2011] are sampled in $E = 6$ levels.

- **Object characterization.** We respect the [Tola et al., 2010] configuration of DAISY and the [Salti et al., 2014] geometrical configuration of SHOT in the construction of **R-DAISY** and **R-SHOT** respectively. For the **R-SHOT** and the **SHOT** descriptors we use an sphere radio of $R = 15\iota_\psi$ to define the support, where ι_ψ is the scale of the described singular-point detection—as described in section 8.4—. For the description of each sphere sector in the **R-SHOT** descriptor, we use the configuration described in 8.4 with the there-mentioned 8x8 bins per two-dimensional histogram, both for the normal and the *Lab* vectors. For the **SHOT** descriptor we use the standard SHOT configuration for the Kinect scenario: 16-bins-histograms per sphere sector to encapsulate the normal information and 5-bins-histograms for the colour information—as reported in [Salti et al., 2014]—.

- **SOM training .**The size HxW of the model (SOM) for each descriptor is a function of the number and the principal values dispersion of the training samples—as described in section 8.5—. The public code available at [Vesanto et al., 2000] was used for creating the SOM. The resulting sizes of the SOM building process for each descriptor as well as their associated mean quantization error (mqe) and mean topology error (mte) are collected in Table 8.1 (see again [Vesanto et al., 2000] for details).

	H	W	mqe	mte
$SOM_{R-DAISY}$	16	5	0.12	0.03
SOM_{R-SHOT}	112	87	0.62	0.04
SOM_{SHOT}	76	58	0.50	0.03

Table 8.1: Details of trained models (SOM) for each description type

Classification statistics.

For evaluation, we have annotated the correct label for each object point in the analysed dataset: thus creating a ground-truth—see bottom-left part of Figure 8.9—.

Let N be the number of objects to identify (i.e. the number of trained classes); let O'_i be the set of test instances for object i, which overall results in N test sets; and let z_j, $j = 1...card(O'_i)$ be the number of object-points of each object instance. This adds up to $Z_i = \sum z_j$ i-object-points: $\overline{x}_{i,\kappa}$, $1 \leqq \kappa \leqq Z_i$, and up to $Z = \sum Z_i$ total object-points to test.

Let $\mathbf{x}_0$ be a specific object-point to which a label $\mathcal{L}(\mathbf{x}_0)$ and associate final score $\Upsilon(\mathbf{x}_0)$ have been obtained for one of the proposed descriptors under one of the three configurations and let $\mathcal{L}_{\mathcal{GT}}(\mathbf{x}_0) = i$ be the manually annotated label for such object-point.

We either classify $\mathbf{x}_0$ as a false negative for object i —$FN_i(\mathbf{x}_0, th) = 1$—if $\Upsilon(\mathbf{x}_0) \leq th$, and $\mathcal{L}(\mathbf{x}_0) = i$, ($FN_i(\mathbf{x}_0, th) = 0$, otherwise); as a true positive for object i—$TP_i(\mathbf{x}_0, th) = 1$— if $\Upsilon(\mathbf{x}_0) > th$ and $\mathcal{L}(\mathbf{x}_0) = i$, ($TP_i(\mathbf{x}_0, th) = 0$, otherwise); or as a false positive for object i' —$FP_{i'}(\mathbf{x}_0, th) = 1$—if $\Upsilon(\mathbf{x}_0) > th$ but $\mathcal{L}(\mathbf{x}_0) = i', i' \neq i$, ($FP_{i'}(\mathbf{x}_0, th) = 0$, otherwise).

The value of th is swapped in the whole range of available scores to measure *Recall* and *Precision* statistics for each object by following their classical equations (see chapter 3) . In particular, to overall evaluate the operation of the descriptor under a given configuration, the global *Recall* and *Precision* statistics can be computed by aggregating results for all the objects.

$$Recall(th) = \frac{\sum_{i=1}^{N} \sum_{\kappa=1}^{Z_i} TP_i(\mathbf{x}_{i,\kappa}, th)}{\sum_{i=1}^{N} \sum_{\kappa=1}^{Z_i} TP_i(\mathbf{x}_{i,\kappa}, th) + \sum_{i=1}^{N} \sum_{\kappa=1}^{Z_i} FN_i(\mathbf{x}_{i,\kappa}, th)} \tag{8.29}$$

and:

$$Precision(th) = \frac{\sum_{i=1}^{N}\sum_{\kappa=1}^{Z_i} TP_i(\mathbf{x}_{i,\kappa}, th)}{\sum_{i=1}^{N}\sum_{\kappa=1}^{Z_i} TP_i(\mathbf{x}_{i,\kappa}, th) + \sum_{i=1}^{N}\sum_{\kappa=1}^{Z_i} FP_i(\mathbf{x}_{i,\kappa}, th)} \tag{8.30}$$

We define th^* as the optimal operation point, i.e. the value of th that produces the closest pair $[Recall(th^*), Precision(th^*)]$ to $[1, 1]$ for each descriptor under each defined configuration. To better asses the overall operation of each descriptor, the $F_1 - score(th^*)$ at the optimal operation point is also computed, following:

$$F_1 - score(th^*) = 2\frac{Precision(th^*)Recall(th^*)}{Precision(th^*) + Recall(th^*)} \tag{8.31}$$

Regarding the calculation of the confusion matrix C_{NxN+1}, we extract its statistics at the optimal operation point th^*. For two different objects i and i' their corresponding position in the confusion matrix $C(i, i')$ contains the proportion of miss-classifications as object i' of object points ground-truth-labelled as object i:

$$C(i, i') = \frac{\sum_{k=1}^{Z_i} FP_{i'}(\mathbf{x}_{i,\kappa}, th^*)}{\sum_{\kappa=1}^{Z_i} TP_i(\mathbf{x}_{i,\kappa}, th^*) + \sum_{\kappa=1}^{Z_i} FN_i(\mathbf{x}_{i,\kappa}, th^*)} \tag{8.32}$$

The diagonal of the confusion matrix $C(i, i)$ contains the proportion of correct classifications of $i-$object points, i.e. the *Recall* at the optimal operation point:

$$C(i, i) = Recall(th^*) \tag{8.33}$$

Finally, for th^*, some object-points may remain unclassified (U). For each object i the proportion of unclassified object-points is equal to the false negative rate at th^*:

$$C(i, U) = \frac{\sum_{\kappa=1}^{Z_i} FN_i(\mathbf{x}_{i,\kappa}, th^*)}{\sum_{\kappa=1}^{Z_i} TP_i(\mathbf{x}_{i,\kappa}, th^*) + \sum_{\kappa=1}^{Z_i} FN_i(\mathbf{x}_{i,\kappa}, th^*)} \tag{8.34}$$

Quantitative results

Tables 8.2,8.3 and 8.4 include the *Precision* , *Recall* and $F_1 - score$ statistics for the R-DAISY, SHOT and R-SHOT descriptors, extracted at the optimal operation point for each object to be identified and for the overall operation of the system (last row of each Table). Additionally, Figures 8.10, 8.11 and 8.12 include—in this order—the confusion matrices obtained for the R-DAISY, SHOT and R-SHOT descriptors. Results are discussed in section 8.7.

Qualitative results

Figure 8.13 includes qualitative results for some of the test frames. These aim to illustrate the operation of the evaluated solutions when facing challenging situations. Results are discussed

and related to quantitative results in section 8.7.

Result's discussion and approach limitations

On a descriptor basis.

The **R-DAISY** descriptor presents, by far, the worst performance among the evaluated. This can be observed in its operation statistics in Tables 8.2,8.3 and 8.4, where even achieving the highest *Recall* rates, these are obtained at very low *Precision* rates, hence leading to very small $F_1 - score$ figures. This behaviour is also clear in the confusion matrices of Figure 8.10. There, the *astronaut* object (and in a lower extent, the *tape* object) appears to concentrate most of the information, and thus, it is the preferred object for most of the test instances when R-DAISY is configured as c.1. or c.2. Whereas configuration c.3 partially solved this issue (there the agglutinating object is *jar*), it leads to almost useless results, as it is clearly illustrated in Figure 8.13. There might be several causes for this operation. R-DAISY is the only approach that does not rely on singular-points for the description, which in turn generates a balancing problem. The amount of training descriptions available is small when compared with the **R-SHOT**, that generates a vector per region and singular-point combination (see for instance the resulting size of $SOM_{R-DAISY}$ in Table 8.1). For this reason, the resulting SOM is small, and of a lower representation capability than the others. As aforementioned, it seems that the label identifying the *astronaut* (followed by the *stapler*) object is the most common among the neurons, as it concentrates the majority of the classifications in c.2. configuration. The high proportion of *astronaut* (8 training instances) and *tape* (7 training instances) bias **R-DAISY** description vectors towards these objects (see configuration c.1. in Figure 8.10). Furthermore, it is worth to mention that **R-DAISY** is the only descriptor amongst the three evaluated that does not handle scale changes—as it does not rely on singular-points—; hence, its lack of invariance to these changes severely harms its operation. Finally, it is interesting to see how distribute encoding changes the picture, leading to a less biased classification system. However, resulting classification approach is still useless for the faced scenario. Probably, both a training with balanced input samples, an increase of the size of the SOM and a proper scheme to handle scale changes would improve the **R-DAISY** descriptor performance.

SHOT descriptor appears to operate with reasonably accuracy in the faced scenario. As aforementioned **SHOT** is the only evaluated descriptor that gives results at object level, i.e. it either classifies or miss-classifies a whole object instance. This identification scheme biases it overall operation to results obtained for large objects. Large objects—due to their distinctive size—are partially isolated from its surrounding objects—e.g. the *jar* or the *carton of juice*—; hence, the majority of their appearance is conserved, non occluded. For the same reason, it also benefits for the isolation of some objects in the test scenes—see *stapler*, *rubik cube* and *glass* in the qualitative results in Figure 8.13—.

Notwithstanding, this holistic operation severely harms the performance of **SHOT** for objects that in test frames are inside the clutter. This is the case for the *cup*, an object which is not detected for the c.1 nor for the c.3 configurations (see 8.11). Furthermore, the use of local reference frames extracted on the whole object apparently harms the operation of **SHOT** for deformable objects as the *tripod* (see Figure 8.13 but also the associated statistics in Tables 8.2,8.3 and 8.4).

R-SHOT descriptor under c.3 is the top performing algorithm amongst the evaluated. **R-SHOT** performs reasonably accurate in all the objects associated with captured—not estimated—depth information. The decrease of performance for these objects can be observed in its operation on crystal (*jar* and *bottle*) and metallic objects (*metal piece*). This suggests that accurate 3D information is required for the **R-SHOT** to achieve high-quality performance (see Tables8.2,8.3 and 8.4).

The same problem can be observed in the confusion matrix of Figure 8.12. A qualitative example of this problem is also observable in Figure 8.13, where the *metal piece* is only fully identified in the last column example and just partially identified in the second column example, whereas the *bottle* is fully identified in the third column example but fails to be identified in the fourth column example—the pure description based comparison (c.1) at least identifies the bottle label—. On contrary, note the excellent behaviour for the *tripod* object, which is highly deformable, clearly benefiting for the part-based modelling. Finally, note how—for the third configuration—the system is able to identify severe occluded objects—which boundaries are highly different than those observed in the training state—, as the *cup* in the third column example or the *spray* in the fourth column example.

On a configuration-basis

Regarding the evaluated configurations, c.2 operates the worse for all the descriptors—in spite of helping in some cases, as in the identification of *tripod* by the **SHOT** description—. In our opinion this is mainly due to the neuron labelling procedure that we have followed (see equation 8.7). As a neuron might be a representative of several classes, all of them should be taken into account when labelling it, and not just the most frequent—i.e. a distribute or weighted labelling procedure may be beneficial—.

Using distributed encoding (c.3) benefits the operation of the descriptors themselves (c.1) not by substantially increasing the per-object performance—which is, instead sometimes decreased—but by balancing the results amongst objects. This can be observed by comparing the first and third confusion matrices in Figures 8.10, 8.11 and 8.12.

In our opinion, this is a strong indicator of the capability of distribute encoding to face unbalanced training, i.e. to put the trained descriptions on even grounds for identification independently of the number of samples used for the training of each object instance.

Overall comparison

In the light of the results, and as aforementioned, the **R-SHOT** descriptor leads the comparison in its c.3 configuration being followed by itself under configuration c.1 and then by **SHOT** under configuration c.3 and **SHOT** in a pure description basis c.1 (see overall $F_1 - score$ row in Table 8.4). **R-SHOT** achieves a 20 % percent of improvement in terms of $F_1 - score$ over the best configuration for SHOT.

On a per object basis, **R-SHOT** also leads results in the classification of ten out of fourteen objects (see Figure 8.12). Anyway, its operation is stable for every other object, but for those which depth information is estimated (not captured) as discussed previously in this section.

The use of the region-based description schemes improve holistic performance in the case of R-SHOT but fails in the case of R-DAISY. However, to be fair, a comparison with unmasked DAISY should be also performed. Nevertheless, this suggest that depth and luminance information by themselves are not enough to provide robust descriptions for identification—when 3D information is available and as in this case, required—.

The part-based modelling generally improves holistic modelling for eleven out of the fourteen tested objects—compare the diagonals of the confusion matrices in the top-row of Figures 8.11 and 8.12—.

The distributed modelling generally improves the operation for all the evaluated descriptors whereas the SOM itself, used as codebook, is not able to adequately discriminate the different object instances.

Approach limitations

The main limitation of the proposed approach is the assumption that objects have been previously segregated. This is in fact a very challenging task, specially for the faced scenarios. One may think that the proposed algorithm puts the cart before the horse, albeit, both its design and our future work plans the evaluation of the algorithm operation in a dual segregation-plus-identification task—which is a common trend in recent object identification and recognition methods Arbelaez et al. [2012]; Girshick et al. [2014]—. As our analysis units are regions, a proper RS of the image would allow to identify both the objects and the parts contours. Hence, for the method to perform decently, it hypothetically would only require the definition of a default object class—in order to assign to this class the untrained object instances—. However, our initial experiments indicate that this approach is infeasible, partially due to the complexity involved in the definition of this *hotchpotch* default class, but mainly due to the multi-coarse RS segmentation scheme proposed in section 8.4. In the proposed dataset, objects are smaller than background and their appearance partially resembles that of the background. If at the output of the multi-coarse segmentation the objects are merged with the background at a particular level of coarseness, the relative information of the objects respect to that of the background is

a minority, and hence, the probability of being an instance of the default class is higher in this level than that of being an object in any other coarseness level. We aim to further inspect this problem and propose solutions to handle it.

8.8 Chapter conclusions.

In this chapter we have presented an approach to object identification partially inspired by psychophysical considerations. The approach relies on a distributed knowledge encoding of the objects parts, these characterised via region-based local features. The system organises evidences from a very small collection of objects—without the requirement of CAD models—in a single and relatively simple neural structure. The approach has been proven to provide promising results in the identification of severely occluded objects by using a very small and short-varied subset of objects in the training stage. In particular, reported results show the benefits of including a region-based version of a recent state-of-the-art 3D descriptor in the proposed neural-framework. Furthermore, the benefits of using distributed encoding in the faced scenarios have been also experimentally evaluated with success.

Precision	**R-DAISY**			**SHOT**			**R-SHOT**		
(th^*)	**c.1**	**c.2**	**c.3**	**c.1**	**c.2**	**c.3**	**c.1**	**c.2**	**c.3**
Astronaut	.05	.05	.05	.23	.00	.24	.21	.16	.26
M. Piece	.12	.12	.12	.29	.00	.46	.38	.10	.53
Bottle	.04	.04	.04	.12	.07	.22	.53	.05	.43
Glass	.12	.12	.12	1.0	.13	1.0	.87	.32	.86
Tape	.11	.11	.11	.54	.26	.42	.52	.43	.45
Jar	.11	.11	.11	1.0	.09	.92	.62	.22	.51
Tripod	.03	.03	.03	.41	.47	.36	.64	.74	.85
Rubik c.	.05	.05	.05	1.0	.14	.53	.87	.14	.77
C. of Juice	.14	.14	.14	.97	.08	.78	.95	.24	.88
Stapler	.04	.04	.04	.86	.00	1.0	.67	.09	.93
Case	.28	.28	.28	1.0	.16	.92	1.0	.46	.82
Spray	.05	.05	.05	.00	.06	1.0	.93	.16	.95
Toy car	.06	.06	.06	.45	.00	.82	.48	.16	.63
Cup	.10	.10	.10	.00	.11	.00	.88	.43	.95
Overall	.07	.07	.07	.38	.13	.52	.60	.26	.64

Table 8.2: Precision statistics at the optimal operation point for the three descriptors under the three configurations analysed. Best operation per object is highlighted in red. See text for c.1, c.2 and c.3 definition.

Recall	R-DAISY			SHOT			R-SHOT		
(th^*)	c.1	c.1	c.1	c.1	c.2	c.3	c.1	c.2	c.3
Astronaut	1.0	1.0	1.0	1.0	.00	.99	1.0	1.0	1.0
M. Piece	1.0	1.0	1.0	.95	.00	.66	.97	1.0	1.0
Bottle	1.0	1.0	1.0	.96	1.0	.93	1.0	.73	.95
Glass	1.0	1.0	1.0	.96	1.0	.94	.91	1.0	1.0
Tape	1.0	1.0	1.0	.96	1.0	.87	.99	.99	.98
Jar	1.0	1.0	1.0	1.0	1.0	1.0	.78	.99	1.0
Tripod	.77	.00	1.0	1.0	1.0	.88	1.0	.99	.95
Rubik c.	1.0	1.0	1.0	.96	1.0	.97	.69	1.0	1.0
C. of Juice	1.0	1.0	.97	1.0	1.0	1.0	1.0	1.0	1.0
Stapler	1.0	1.0	1.0	1.0	.00	.93	.98	1.0	.99
Case	1.0	1.0	1.0	1.0	1.0	.93	.90	1.0	1.0
Spray	.96	.96	1.0	.00	1.0	.88	.91	1.0	1.0
Toy car	1.0	1.0	1.0	1.0	1.0	.89	1.0	1.0	1.0
Cup	1.0	1.0	1.0	.00	1.0	.00	.99	.94	.97
Overall	.99	.99	1.0	.97	.99	.88	.94	.99	.99

Table 8.3: Recall statistics at the optimal operation point for the three descriptors under the three configurations analysed. Best operation per object is highlighted in red. See text for c.1, c.2 and c.3 definition.

F-Score	R-DAISY			SHOT			R-SHOT		
(th^*)	c.1	c.1	c.1	c.1	c.2	c.3	c.1	c.2	c.3
Astronaut	.10	.11	.14	.38	-	.39	.35	.27	.41
M. Piece	.22	.16	.20	.45	-	.54	.55	.17	.69
Bottle	.07	.10	.13	.22	.13	.36	.69	.09	.59
Glass	.22	.23	.20	.98	.23	.97	.89	.49	.92
Tape	.20	.13	.10	.69	.42	.56	.68	.60	.62
Jar	.20	.08	.16	1.0	.17	.96	.69	.36	.68
Tripod	.06	-	.13	.58	.64	.51	.78	.84	.90
Rubik c.	.09	.08	.05	.98	.25	.68	.77	.24	.87
C. of Juice	.25	.27	.34	.98	.14	.88	.98	.39	.93
Stapler	.08	.08	.10	.93	-	.97	.80	.16	.96
Case	.43	.22	.12	.10	.27	.92	.95	.63	.90
Spray	.10	.10	.04	-	.12	.94	.92	.28	.97
Toy car	.11	.13	.06	.62	-	.85	.65	.28	.77
Cup	.18	.36	.06	-	.21	-	.93	.59	.96
Overall	.14	.13	.13	.55	.23	.65	.73	.41	.78

Table 8.4: F-Score statistics at the optimal operation point for the three descriptors under the three configurations analysed. Best operation per object is highlighted in red. See text for c.1, c.2 and c.3 definition.

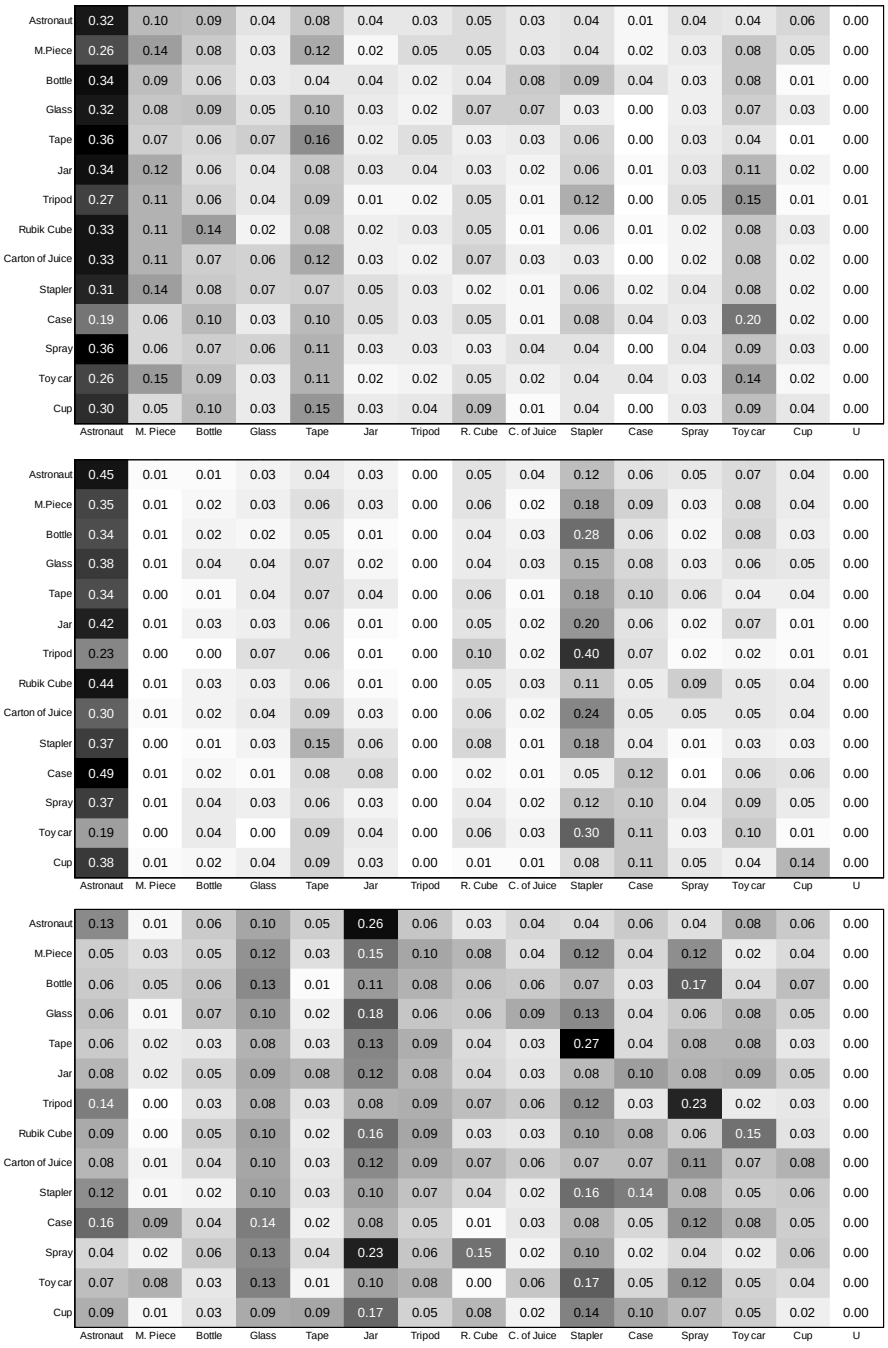


Fig. 8.10. . Confusion matrices for R-DAISY descriptor under configurations: c.1 (first row), c.2 (second row) and c.3 (third row).The wither (blacker) the lower (higher). Best operation would be a full black diagonal on a white background. Rows are tested, columns detected.

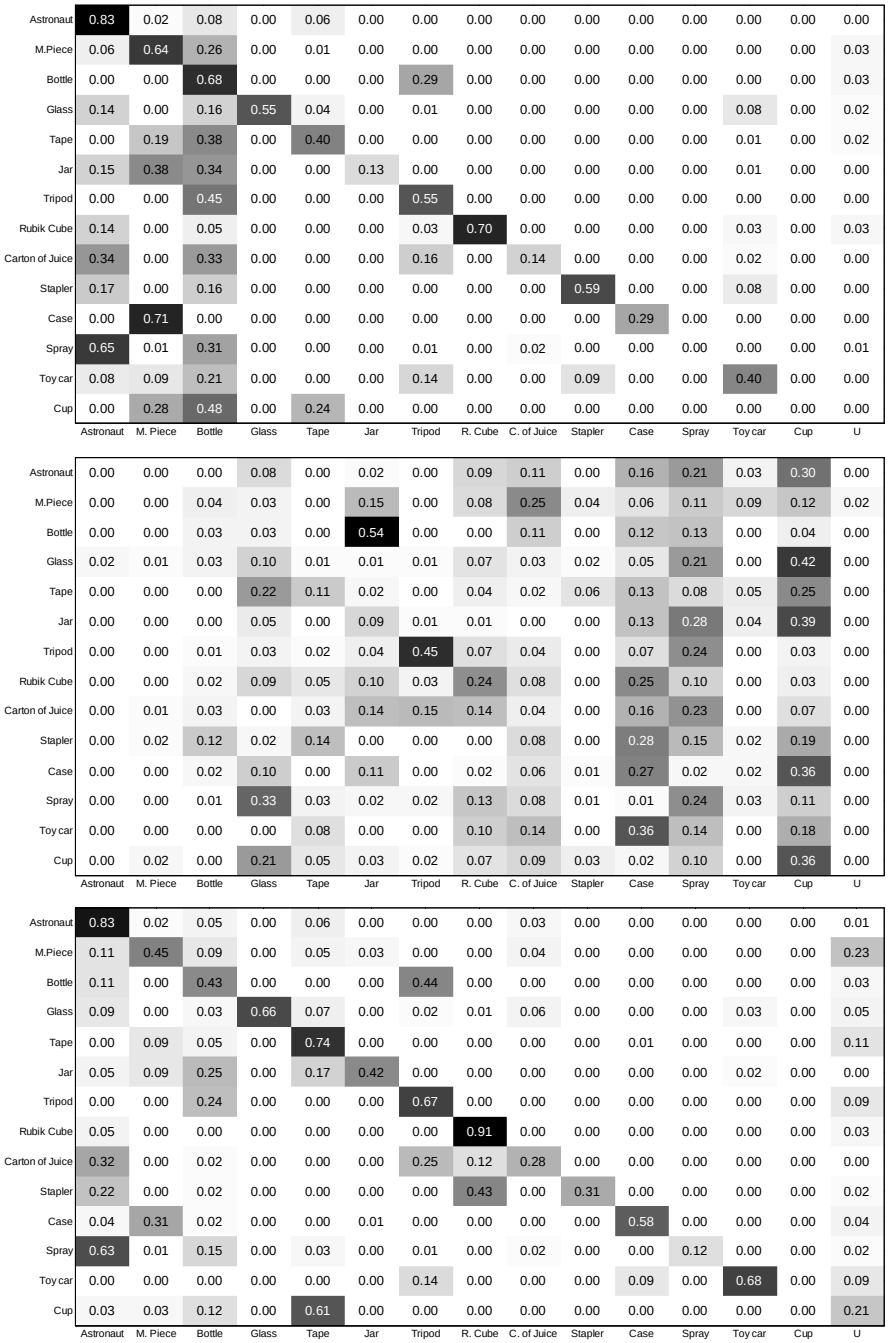


Fig. 8.11. . Confusion matrices for SHOT descriptor under configurations: c.1 (first row), c.2 (second row) and c.3 (third row). The wither (blacker) the lower (higher). Best operation would be a full black diagonal on a white background. Rows are tested, columns detected.

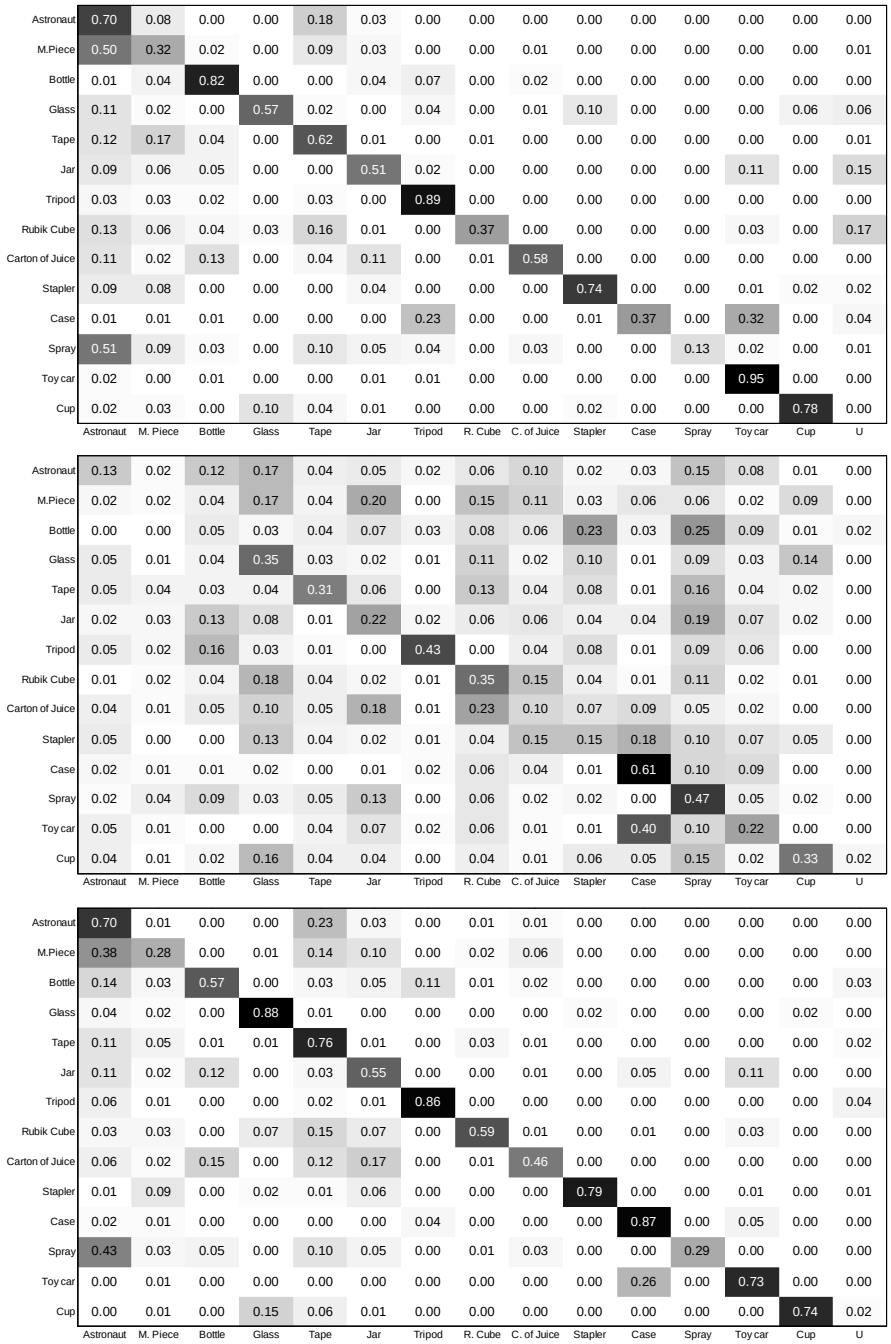


Fig. 8.12. . Confusion matrices for R-SHOT descriptor under configurations: c.1 (first row), c.2 (second row) and c.3 (third row).The wither (blacker) the lower (higher). Best operation would be a full black diagonal on a white background. Rows are tested, columns detected.

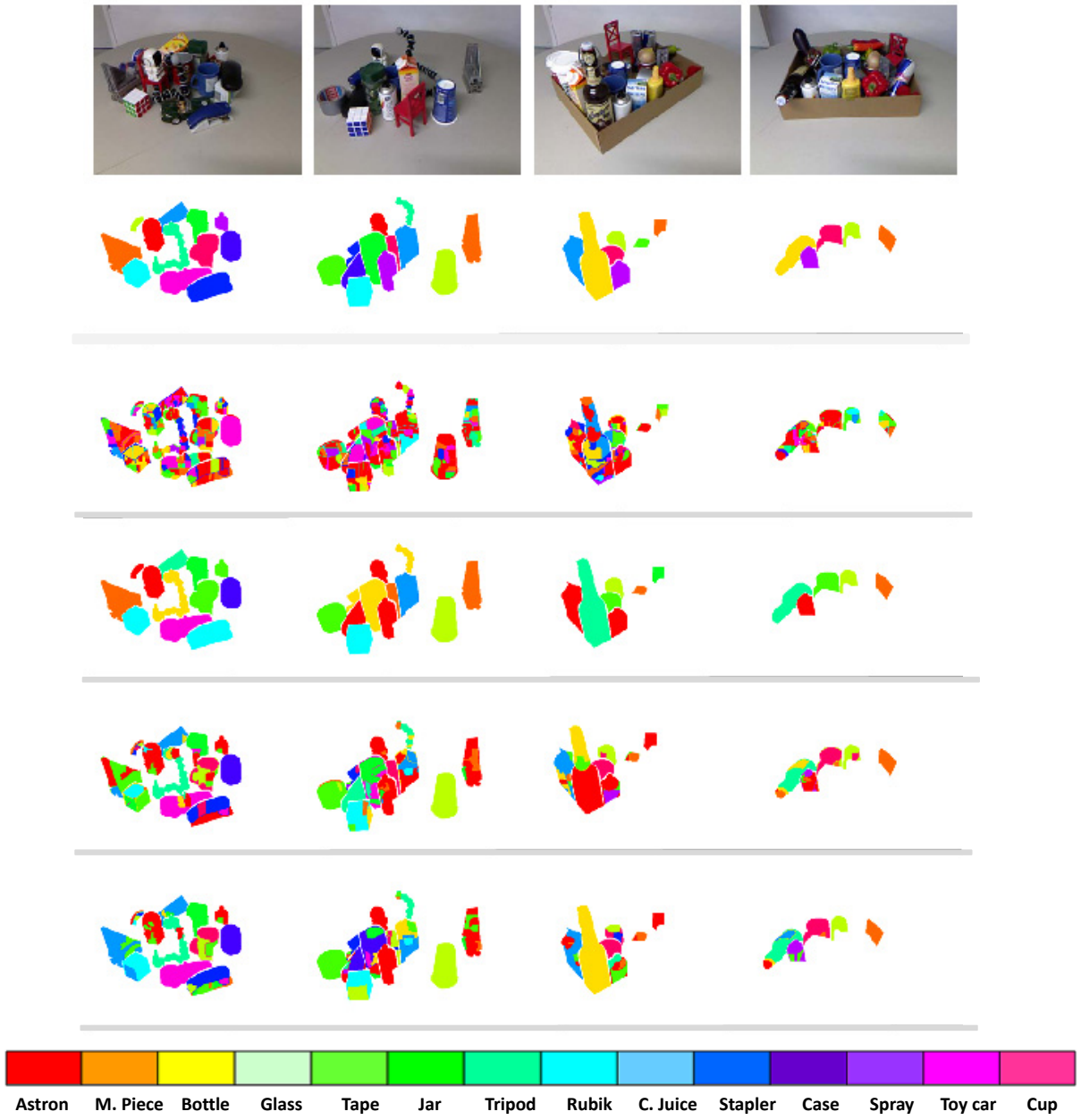


Fig. 8.13. Qualitative results. Test images (first row) and associated ground-truth (second row). Results for R-DAISY descriptor configured as c.3 (third row), for the SHOT descriptor under c.3 configuration (fourth row), for the R-SHOT descriptor under c.1 (fifth row) and for the R-SHOT descriptor configured as c.3 (sixth row). Associated labels are included in the bottom row of the Figure to ease visualization. Note that result images have been recomposed as each object is analysed isolated.

Chapter 9

Projective deformation and appearance transformation of region-supports for wide-baseline point matching.

9.1 Wide-baseline point correspondences: benefits and challenges.

The ability to match image points across images is a key stage in several tasks of computer vision, including: simultaneous localization and mapping (Fuentes-Pacheco et al. [2015]), image mosaicking (Joshi and KhomL alSinha [2013]), image-based rendering (Zhang and Chen [2004]) and object detection and recognition (Xu and Zhang [2013]). The point correspondence problem can be defined as the task of establishing unequivocal relationships between image points given at least two—different—views of a scene.

For a given point, the task can be mainly described by a two-stage procedure: point description and point searching. The description stage compiles the definition of the support or description area around the point and the feature extraction process. In the searching stage, this description is transformed and adapted under a set of constraints and then compared to point descriptions in the other image.

According to the separation between each pair of neighbouring views, two scenarios might be distinguished: small- and wide-baseline. The small-baseline scenario encompasses well-founded tasks in computer vision including: stereo-vision (Scharstein and Szeliski [2002]) and optical flow estimation (Weinzaepfel et al. [2013]). In these scenarios, strong geometry constraints on the expected position of the image point on the views can be imposed. On the contrary, captured

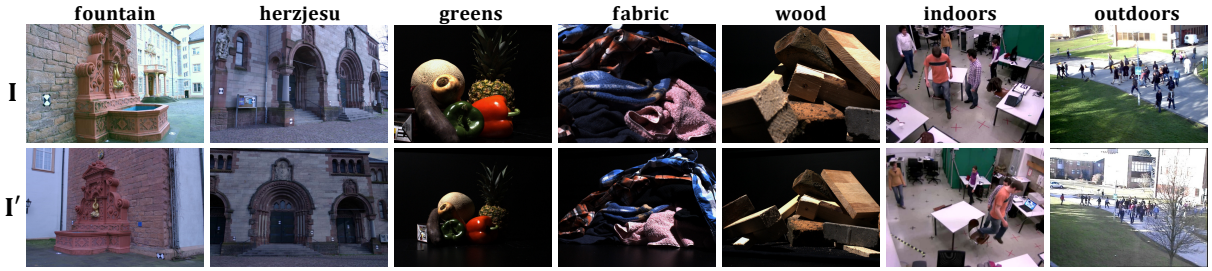


Fig. 9.1. The data set analysed in the experiments, including wide-baseline typical challenges: strong projective deformations, global and local illumination changes and inter-object occlusions, as well as multiple oriented surfaces. Images are extracted from previous data sets: Strecha et al. [2008] (fountain and herzjesu), Aanaes et al. [2012] (greens, fabric and wood), Possegger et al. [2013] (indoors) and Ferryman et al. [2009] (outdoors).

images in wide-baseline scenarios are subject to large projective distortions, intense appearance changes and severe occlusions (Mikolajczyk et al. [2005]; Yu and Morel [2009]; Tola et al. [2010]). In Aanaes et al. [2012], experiments shown how large changes in the capture point-of-view and captured illumination severely affect the performance of existing correspondence searching methods. Figure 9.1 qualitatively describes these challenges by exemplifying the target scenarios of the proposed approach.

State-of-the-art solutions to these challenges can be roughly divided into two branches: invariance and adaptation. The former includes those methods which design robust description schemes able to cope with the aforementioned challenges, whereas approaches in the latter adapt the descriptions by inferring the appearance and geometric transformations to which the point is subject in the other view. The solution proposed in this chapter lies in between these two branches: occlusions are considered in the description stage and projective and appearance transformations are corrected in the searching stage. In the core of this hybrid solution lies the strategy to obtain description supports as bigger and as descriptive as possible.

In order to simplify the matching problem, the scene can be conceptually approximated as piecewise planar. From this approximation emerges a plane-based strategy to establish correspondences. An image point $\boldsymbol{\psi}$ is the projection of a scene point $\boldsymbol{\Psi}$. This scene point lies on a surface, which under the assumption, can be approximated by a plane. This plane is also projected onto the images, and these plane projections contain the projections of the point $\boldsymbol{\Psi}$: $\boldsymbol{\psi}$ and $\boldsymbol{\psi}'$—the corresponding point to $\boldsymbol{\psi}$ in another image—. Hence, $\boldsymbol{\psi}'$ can be located by relating the projections of the plane surface on which the scene point lies.

Under this simplification, state-of-the-art descriptors are designed under the assumption that projective transformations of the scene can be modelled as locally affine; hence deformations of the plane projections—on which the description support is assumed to be confined—are a function of just 6-parameters. This assumption is only valid for very small supports. In contrast,

our aim is to correct the projective distortions themselves, increasing the complexity—up to 8-parameters—, but allowing the use of bigger supports, therefore, including more image evidences in the point description. We are not the first to propose this scheme; it constitutes the basis of plane-sweeping approaches (Friedrich Fraundorfer [2006]; Micusik [2009]). However, our method is, to our knowledge, the first able to face scenarios where no specific constraints on the orientations of the scene surfaces are imposed.

In the approach proposed in this chapter—a flowchart is depicted in Figure 9.2—we propose a description and matching method, assuming that the points of interest have been detected or that the detection stage is unnecessary (as in dense approximations). An image point—from now on, the anchor point—will be described by the spectral properties of a neighbourhood around it (in the feature extraction module). A pre-detection of the image edges (edge extraction module) on the source image is used to define a set of weights around the anchor point. These weights measure the resemblance of the anchor point to its neighbourhood (in the weighting-by-resemblance module). For each anchor point, a description area or support around it is automatically defined by measuring the sparsity of the edge distribution in the point neighbourhood (in the support definition module). The image points in this support constitute the description locations or description samples of the anchor point. The anchor point location on the image lattice and the scene calibration are used for constraining the searching space in the other image (in the hypotheses generation module), and to define a set of possible geometrical transformations of the anchor point support. Projective distortions are then considered by searching on this hypothesised subset of transformations of the support (in the support transformation module), by means the use of plane-induced homographies (in the geometric transformation module). Under each of these transformation hypotheses, a possible location and geometric configuration of the support description samples on the other image is defined. Features are then extracted on the other image on the samples defined by these projected locations (feature selection). In order to also cope with appearance changes, a self-adaptive feature transformation model is designed. The contribution of each description sample in the transformation model is higher the higher is its resemblance—codified by the weights in the weighting-by-resemblance module—to the anchor point (in the feature transformation module). Finally, the goodness of each hypothesis is measured by comparing the support of the anchor point and each hypothesised projected support by means of the transformed features (in the hypothesis scoring module). Again, the influence of each description sample in the comparison is function of its resemblance (relative weight) to the anchor point.

The rest of the chapter is organized as follows. Related work is discussed in section 9.2. Sections 9.4 and 9.5 describe the proposed description and searching schemes respectively. The solutions there presented are evaluated in section 9.6 leading to a set of conclusions and inspiring the future work which are described in section 9.7.

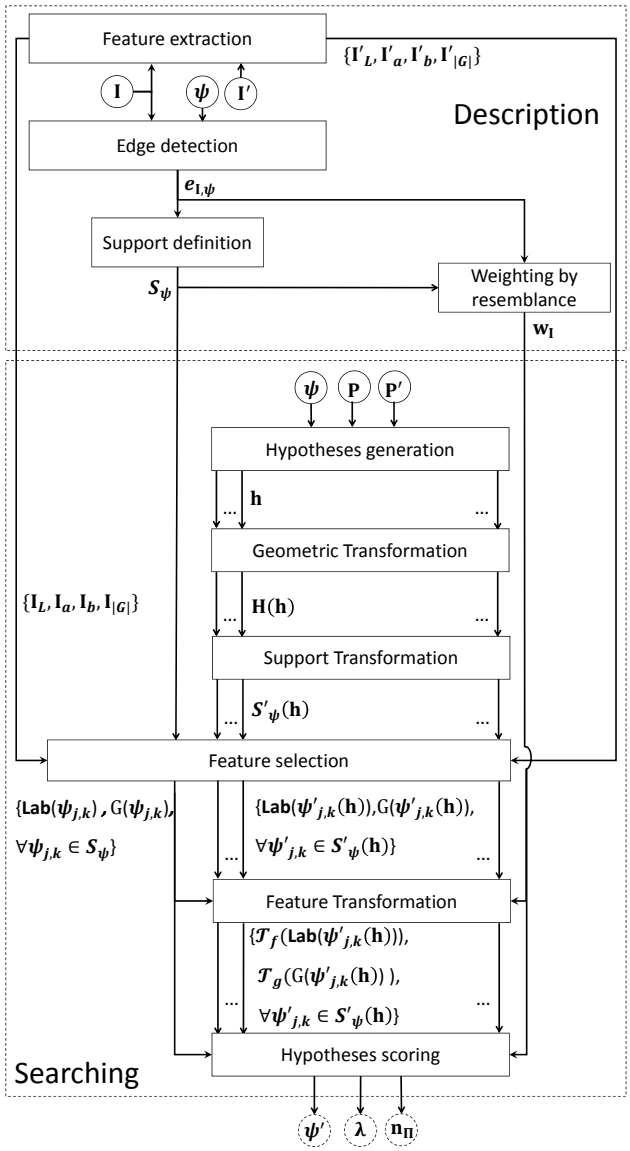


Fig. 9.2. Flowchart of our method. Inputs for the algorithm (solid-line circumferences) are: two scene views ($\mathbf{I}$ and $\mathbf{I}'$), their associated calibration ($\mathbf{P}$ and $\mathbf{P}'$) and an anchor point. The output for an anchor point ψ on $\mathbf{I}$ are (dashed-line circumferences) estimations of its corresponding point ψ' on $\mathbf{I}'$, its distance λ to the camera that captures $\mathbf{I}$ and of the normal vector $\boldsymbol{n}_{\boldsymbol{\Pi}}$ of the surface on which its 3-dimensional projection lies. See text for details and Table 9.1 for symbol definitions.

9.2 Invariant vs adaptable descriptions.

Table 9.2 compares top-referenced recent approaches to point matching in wide-baseline scenarios. It should be noted that the comparison is performed on a design-basis with independence

Symbol	Type	Description
$\mathbf{I}$,$\mathbf{I}$'	RGB-images	source and reference images
$\mathbf{P}$,$\mathbf{P}$'	$3x4-$matrix	camera projection matrices
$\boldsymbol{\psi}$,$\boldsymbol{\psi'}$	2-d point in homogeneous	image points on $\mathbf{I}$ and $\mathbf{I}$'
$\boldsymbol{\Psi}$	3-d point in homogeneous	scene point
λ	scalar	depth for a scene point
λ_{Π}	scalar	depth for a scene plane
$\boldsymbol{n_{\Pi}}$	vector	normal for a scene plane
$(\mathbf{h})$	set	hypothesis set (depth and normal)
$\mathbf{H}(\lambda_{\Pi}, \boldsymbol{n_{\Pi}})$	$3x3-$matrix	plane induced homography
$\boldsymbol{S_{\psi}}$	set of image points	description support (set of description samples)
$\mathbf{Lab}(\boldsymbol{\psi})$	vector	CIE-Lab descriptor
$G(\boldsymbol{\psi})$	scalar	gradient descriptor
$e_{\mathbf{I},\boldsymbol{\psi}}$	scalar	intensity of the edge respect to $\boldsymbol{\psi}$ on $\mathbf{I}$
$\mathbf{w_I}$	vector	weight (resemblance) vector
$\mathcal{T}_g$,$\mathcal{T}_f$	functions	appearance transformation functions

Table 9.1: Table of main symbols used along the chapter.

of their experimental results. As aforementioned, these approaches can be divided into *invariant*, invariance branch, and *adaptable*, adaptation branch, according to their strategy to face the considered challenges. As this chapter presents a point matching method, we only focus in the description and matching part of these approaches, leaving the detection out of the comparison.

The comparison is performed on different categories. The description strategy of each approach is indicated in terms of nature and shape of the description support and in terms of the features used for description. Then, robustness of each approach to wide-baseline challenges—occlusions, appearance changes of different nature and geometrical deformations—is analysed. Approaches are either claimed to be designed (✓) or not-designed (empty) to cope with these challenges. In order to provide additional information about the discussed solutions—not just a binary classification—, their specific particularities of the methods are indicated (see caption of Table 9.2). Next subsections describe and compare the methods included in Table 9.2.

	Approaches	Description scheme			Occlusion	Appearance changes			Transformations					Prior
		support	size	features	Handling	Y (gain)	Y (bias)	Chroma	Isometric	Similarity	Affinity	Projective	Non-rigid	
invariant	SIFT Lowe [1999]	patch	small	HOG		✓			✓	$\checkmark^{d.}$				
	Daisy Tola et al. [2010]	polar	small	HOG	$\checkmark^{p.}$	✓			✓					F
	LIOPZhenhua Wang and Wu [2011]	circular	small	L		✓			✓	$\checkmark^{d.}$				
	SID Kokkinos and Yuille [2008]	log-polar	large	FT		✓	✓		✓	✓				
	S-SID Trulls et al. [2013]	log-polar	large	FT	✓	✓	✓		✓	✓				
	DaLI Simo-Serra et al. [2015]	mesh	medium	HKS		✓	✓		$\checkmark^{c.}$	$\checkmark^{c.}$			✓	
adaptable	GPM Barnes et al. [2010]	patch	small	L					✓	✓				
	NRDC HaCohen et al. [2011]	patch	small	Lab+G		✓	✓	✓	✓	✓				
	A-SIFT Yu and Morel [2009]	patch	small	HOG		✓			✓	✓	✓			
	ISIFT Yu et al. [2012]	patch	small	HOG		✓	✓		$\checkmark^{1.}$	$\checkmark^{1.}$	$\checkmark^{1.}$	$\checkmark^{1.}$		H
	Plane-Sweep Micusik [2010]	region	region	RGB+G			✓		$\checkmark^{3.}$	$\checkmark^{3.}$	$\checkmark^{3.}$	$\checkmark^{3.}$		P
	our proposal	polar	adaptable	Lab+G	✓	✓	✓	✓	$\checkmark^{N.}$	$\checkmark^{N.}$	$\checkmark^{N.}$	$\checkmark^{N.}$		P

Table 9.2: Recent approaches for point-matching in wide-baseline scenarios. $\checkmark^{p.}$ stands for predefined, $\checkmark^{d.}$ and $\checkmark^{c.}$ for provided in the detection and comparison stages and $\checkmark^{1.}$, $\checkmark^{3.}$, $\checkmark^{N.}$ indicate a restriction on the number of orientation surfaces—one, three or multiple—. Prior stands for prerequisites on calibration information: full (P), the fundamental matrix (F) or a plane homography (H).

Invariant descriptions

The SIFT (Scale Invariant Feature Transform) description Lowe [1999] has been experimentally proven to cope with affine distortions to some extent (Mikolajczyk et al. [2005]). However, it is designed to be robust to similarities—4 parameters: support translations, rotations and scale changes—but not to affinities, which also include non-isotropic scaling Yu and Morel [2009]. Instead, detection information is used to compute descriptions at the scale (Lindeberg [1993]) where the point is most prominent. The SIFT descriptor has been improved, either for computational efficiency, as in SURF (Speeded Up Robust Features, Bay et al. [2006]), or by using alternative supports as in GLOH (Gradient Location-Orientation Histogram, Mikolajczyk et al. [2005]). The HOG (Histogram of Oriented Gradients) normalization of the SIFT descriptor has been experimentally proven to cope with moderate monotonic illumination changes. However, the LIOP (Local Intensity Order Pattern) descriptor, explicitly designed to face these challenges, is supposed to provide better results (Zhenhua Wang and Wu [2011]).

The Scale Invariant Descriptor (SID) Kokkinos and Yuille [2008] replaces scale-selection with the Fourier Transform modulus of log-polar transformed supports. Points are described by the first coefficients of the transform (FT in Table 9.2). This strategy also conveys invariance to in-plane rotations. SID has been proven to be robust to up to a 4 order scale change. However, for the FT to be representative, support information should be plenty. To this aim SID includes a big amount of contextual information in the description. This is achieved by the processing of the data at high scales. The effective area covered by the support is substantially increased by this scheme, including image samples not related—being projection of a different scene

surface—with the anchor point. Hence, the influence of these uninteresting or noise surrounding image samples—known as occlusion samples—in the SID description is not only significant, but motivated by its design.

In order to reduce the influence of occlusion samples, the segmentation-aware SID (S-SID) is proposed in Trulls et al. [2013]. Their solution builds upon the strategy first defined in DAISY Tola et al. [2010] of using binary patterns to inhibit occlusion samples in the support. In DAISY, a set of predefined patterns is evaluated and the optimal pattern is selected through a maximum a posteriori (MAP) criterion on the final score. In superpixel SIFT (SP-SIFT) Navarro et al. [2014] these binary patterns are instead obtained automatically from a preliminary region segmentation of the image. Returning to the S-SID approach, the binary nature of the patterns is replaced by a weighted scheme: the contribution of each description sample is measured as a function of the resemblance of the description sample to the anchor point. This last solution—of a fuzzy inspiration—, allows to increase significantly the size of the support. Finally, DaLi (Deformable and Light invariant descriptors) Simo-Serra et al. [2015] motivates the use of Heat Kernel Signatures (HKS) as these represent intrinsic shape descriptors which, by definition, are robust to non-rigid deformations. However, DaLi is not explicitly designed to be robust to rigid transformations and these are faced—to some extent—in the comparison stage.

Adaptable descriptions

The main limitation of existing invariant methods is that all of them are extracted using 2-dimensional kernels on the images domain, hence introducing a bias for scene surfaces parallel to the images planes (fronto-parallel surfaces) . In wide-baseline scenarios, the projections of a scene surface may be subject to large projective distortions, including 3-dimensional rotations of the surface itself. Therefore, the captured appearance may vary significantly from one view to another. In order to face this challenge, adaptation methods aim to infer the surface transformation. Then, the description support is adapted according to such inference.

The GPM (Generalized Patch Match) algorithm (Barnes et al. [2010]) searches for the best patch-to-patch correspondence between two images. The algorithm does not rely in any previous calibration of the cameras; it handles similarities by sweeping over a range of translation, rotations and scaling parameters. The features and metric there used are arbitrary, but mainly focused on the luminance channel. The NRDC (Non-Rigid Dense Correspondence) algorithm HaCohen et al. [2011] extends GPM by considering variations of luminance (gain and bias), colour (bias) and gradient (gain) channels. The swap is performed on seven searching variables—the four related to the features plus translation, rotation and scale—.

ASIFT (Affine-SIFT) Yu and Morel [2009] starts from the invariance of SIFT to similarities and extends it to affinities by sweeping on the two remaining parameters. To this aim, ASIFT proceeds by simulating a discrete set of the plausible affine distortions caused by the change

of the camera optical axis orientation with respect to a fronto-parallel surface. This sweeping scheme allows ASIFT to be autonomous, i.e. it does not require any previous calibration of the scenario. However, ASIFT does not account for feature correction and still relies in invariant methods to face similarities. Alternatively, if calibration information is available or can be estimated, it can be used to constraining the search. Iterative SIFT (ISIFT) Yu et al. [2012] starts from an initial estimation of a plane-to-plane homography given by invariant methods. This initial estimate is refined by luminance correction of the plane in the other image. The algorithm converges to good and numerous correspondences on the plane if the initial estimate is accurate enough.

Plane-sweeping approaches extend this behaviour by testing a family of 3-dimensional plane positions and orientations. The approach can also be extended to scenarios with a higher number of cameras as images do not need to be rectified. In fact, description supports are rectified instead. Plane-sweeping approaches rely on the definition of 3-dimensional plane. A plane is defined by its the normal vector to its surface. This normal is relative to the viewing angle. If the normal is referenced to a given camera, the plane can be described by the norm of the normal vector (which in turn is related with the distance, or relative position, of the plane to the camera) and with the unitary normal vector (which defines the 3-dimensional plane orientation). In the unit sphere, the unitary normal vector is just defined by two angles: azimuth and elevation.

The first plane-sweeping approach dates back to 1996 (Collins [1996]); there, only image-parallel planes are considered and captured projective image distortions are ignored; hence not searching for the unitary plane normal. In Gallup et al. [2007] the three mayor orientations of the scene surfaces are detected by invariant methods (just three unitary normals are considered). These orientations are used to drive sweeping processes on the hypothetical plane position. This results in a family of hypotheses able to cope with urban scenarios, as in these scenarios surfaces are mainly restricted to these three orthogonal orientations. These three orientations coincide with the vanishing directions of a urban scene, thereby this scheme is usually claimed to operate on a *Manhattan world*. In Micusik [2010, 2009] this idea is adapted to the use of super-pixels as analysis units. Super-pixels are claimed to reduce ambiguities in low-textured areas, provide supports adapted to image content and reduce computational cost of post-processing. Plane-sweeping approaches are extended in Schindler and Dellaert [2004], by increasing the number of considered plane orientations to multiple pairs of horizontal vanishing directions, all of them orthogonal to the vertical mayor orientation of a urban scene; hence, the unitary normal vectors of each plane are just defined by one of the angles, as the other is fixed. The mayor orientation is typically set by the identification of the ground plane in the scene. Such scenario is known as the *Atlanta world*. Finally, the MMF (Mixture of *Manhattan* frames) model (Straub et al. [2014]) smooths the urban constraint by analysing surface orientations extracted from the scene itself

(without constraining the unitary normal angles). However, these orientations are obtained by the inference of the surface normals—which, as aforementioned define surfaces orientation—on the scene 3-dimensional point cloud. Unfortunately, such information is only available if the scene is recorded indoors with a Time-of-flight camera (as the Kinect) or outdoors via the LiDAR technology.

Proposed description and point matching approach

In the light of the previous discussions and in order to establish the links and differences of the proposed approach and the state-of-the art we here expose the novel contributions of our proposal:

- In comparison to existing approaches, our proposal is—to our knowledge—the first that considers orientation-unconstrained projective deformations of the description support without the requirement of any 3-dimensional knowledge of the scene—but requiring camera calibration—.
- The proposed method builds upon plane-sweeping approaches by exploring its operation on an orientation-unconstrained scenario. In our opinion, such a scenario was previously ignored mainly due to the complexity inherent to the analysis of the whole range of surface orientations. In this chapter, we show how the hypotheses space can be strongly constrained if the scene calibration is known.
- We propose a scheme for automatic support-size definition. In particular the extend of the plane on which anchor point lies is detected by measuring the sparsity of the response of a fuzzy region segmentation technique.
- We use the resemblance between the anchor point and the description samples used not only to shun the influence of the background and occluding objects in the anchor point description—as in Trulls et al. [2013]— but also to drive a surface-aware appearance transformation model.

9.3 Background: Epipolar geometry and homographies

Depth sweeping

Let $\mathbf{I}$ be an image captured by camera $\mathbf{C}$. Image points in $\mathbf{I}$ can be represented by homogeneous coordinates $\boldsymbol{\psi} = (\alpha u, \alpha v, \alpha)^T, \alpha \neq 0$ which are the result of projecting 3-dimensional points $\boldsymbol{\Psi} = (\beta U, \beta V, \beta W, \beta)^T, \beta \neq 0$ over the $\mathbf{I}$-plane through the line that joints $\boldsymbol{\Psi}$ and the camera optic centre via the $3x4$ camera projection matrix $\mathbf{P} = \beta\mathbf{K}[\mathbf{R}|\mathbf{t}]$. This matrix can be parametrized by

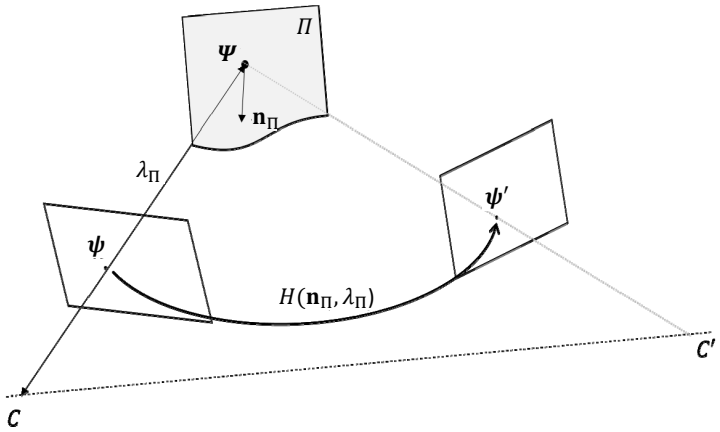


Fig. 9.3. Plane-induced homography. A scene plane $\Pi = (\boldsymbol{n_\Pi}^T, \lambda_\Pi)^T$ defines a plane-induced homography that maps projected points of the scene plane in one image onto the corresponding points in another image

its intrinsic $\mathbf{K}$ and extrinsic parameters: rotation matrix $\mathbf{R}$ and translation vector $\mathbf{t}$ respect to a reference world centre. From now on, we assume that $\beta = 1$, then the camera is normalised and $\boldsymbol{\Psi} = (U, V, W, 1)^T$. The projection of $\boldsymbol{\Psi}$ on $\boldsymbol{\psi}$ is thus defined, under a pin-hole camera model, by:

$$\lambda\boldsymbol{\psi} \coloneqq \mathbf{K}[\mathbf{R}|\mathbf{t}]\boldsymbol{\Psi} \tag{9.1}$$

, where λ is the distance of $\boldsymbol{\Psi}$ to the focal or image plane of $\mathbf{C}$. For the shake of simplicity, λ would be also called depth in some parts of the chapter.

The line that joints $\boldsymbol{\psi}$ with the centre of $\mathbf{C}$ and extends to $\boldsymbol{\Psi}$ is known as the optical ray, which defines the set of 3-dimensional points $\boldsymbol{\Psi}(\lambda)$, that project onto $\boldsymbol{\psi}$:

$$\boldsymbol{\Psi}(\lambda) \coloneqq \lambda \begin{pmatrix} (\mathbf{KR})^{-1}\boldsymbol{\psi} \\ 0 \end{pmatrix} + \begin{pmatrix} -(\mathbf{KR})^{-1}(\mathbf{Kt}) \\ 1 \end{pmatrix} \tag{9.2}$$

Let $\mathbf{I}'$ be another image of the same scene, captured by a camera $\mathbf{C}'$ characterised by the normalised projection matrix $\mathbf{P}' = \mathbf{K}'[\mathbf{R}'|\mathbf{t}']$. Inserting the result of equation 9.2 into the equivalent of equation 9.1 for this camera we obtain:

$$\lambda'(\lambda)\boldsymbol{\psi}'(\lambda) \coloneqq \mathbf{K}'[\mathbf{R}'|\mathbf{t}']\boldsymbol{\Psi}(\lambda) \tag{9.3}$$

, where $\boldsymbol{\psi}'(\lambda)$ is the projection of $\boldsymbol{\Psi}(\lambda)$ on $\mathbf{I}'$ assuming it is at a distance λ of the focal plane of $\mathbf{C}$, whereas $\lambda'(\lambda)$ is the distance of $\boldsymbol{\Psi}$ to the focal plane of $\mathbf{C}'$ under the same assumption. Equation 9.3 defines a process to locate the projection of an image point $\boldsymbol{\psi}$ in $\mathbf{I}$ on $\mathbf{I}'$ under the hypothesis that $\boldsymbol{\Psi}$ is at a depth λ. This is the process known as depth sweeping.

A two dimensional version of this process relies on the projection of the optical ray onto $\mathbf{I}'$. This projection, an epipolar line $\boldsymbol{l}_{\boldsymbol{\psi}}$, constrains the possible matchings in $\mathbf{I}'$ of the point $\boldsymbol{\psi}$ in $\mathbf{I}$. This line can be computed for each $\boldsymbol{\psi}$ by $\boldsymbol{l}_{\boldsymbol{\psi}} = \mathbf{F}\boldsymbol{\psi}$, where $\mathbf{F}$ stands for the $3x3$ fundamental matrix relating the projections captured by $\mathbf{C}$ and $\mathbf{C}'$. $\mathbf{F}$ can be obtained from $\mathbf{P}$ and $\mathbf{P}'$ as described in Hartley and Zisserman [2004].

Plane-induced homographies

Let us consider a scene plane $\Pi = (\boldsymbol{n_\Pi}^T, \lambda_\Pi)^T$, with $\boldsymbol{n_\Pi}$ the plane normal vector which defines its orientation and λ_Π the distance of the plane to the focal or image plane of $\mathbf{C}$ in the direction of $\boldsymbol{n_\Pi}$. The relationship between this scene plane and the image plane is defined by an homography (Hartley and Zisserman [2004]). Furthermore, the relationship between the $\mathbf{C}$ image plane and the $\mathbf{C}'$ image plane is also defined by an homography that maps the projected points of the scene plane Π in the $\mathbf{C}$ image plane onto the corresponding points in the $\mathbf{C}'$ image plane (Hartley and Zisserman [2004]). This is the so-called plane-induced homography.

$$\mathbf{H}(\lambda_\Pi, \boldsymbol{n_\Pi}) \coloneqq \mathbf{K}'(\mathbf{R}'\mathbf{R}^T + \mathbf{R}' \left[\frac{\mathbf{t}' - \mathbf{t}}{\lambda_\Pi} \boldsymbol{n_\Pi}^T \right] \mathbf{R}^T)\mathbf{K}^{-1} \tag{9.4}$$

Note that, if the scene plane is unknown—which is the faced problem—equation 9.4 is a family of homographies with three degrees of freedom or parameters: $\boldsymbol{n_\Pi}^T/\lambda_\Pi$, with $\boldsymbol{n_\Pi}$ the plane normal vector of unit length (two angles) and λ_Π related with the distance of the scene plane to $\mathbf{C}$ (a scalar). These homographies map corresponding image points $\boldsymbol{\psi}$ and $\boldsymbol{\psi}'(\lambda)$ which are the projection of scene points $\boldsymbol{\Psi}(\lambda)$ which lie on each hypothesised plane$\Pi = (\boldsymbol{n_\Pi}^T, \lambda_\Pi)^T$:

$$\boldsymbol{\psi}'(\lambda) \coloneqq \mathbf{H}(\lambda_\Pi, \boldsymbol{n_\Pi}) \times \boldsymbol{\psi} \tag{9.5}$$

, where $\lambda = \lambda_\Pi$ for fronto-parallel planes, can be computed otherwise by the intersection of the optical ray of equation 9.2 with the 3-dimensional plane Π. Due to the compatibility between plane-induced homographies and epipolar geometry, the projected point $\boldsymbol{\psi}'(\lambda)$ always lies on $\boldsymbol{l}_{\boldsymbol{\psi}}$ with independence of the values of λ_Π and $\boldsymbol{n_\Pi}$. The concept of plane-induced homography is illustrated in Figure 9.3.

9.4 Point description

The characterization stage of the proposed approach basically combines the features of NRDC (HaCohen et al. [2011]) with the weighting-by-resemblance scheme of S-SID (Trulls et al. [2013]) under the polar organization of DAISY (Tola et al. [2010]).

In state-of-the-art invariant methods, the size of the support is defined by a scale-space analysis. This scheme returns, for each point, the scale at which a particular feature in the

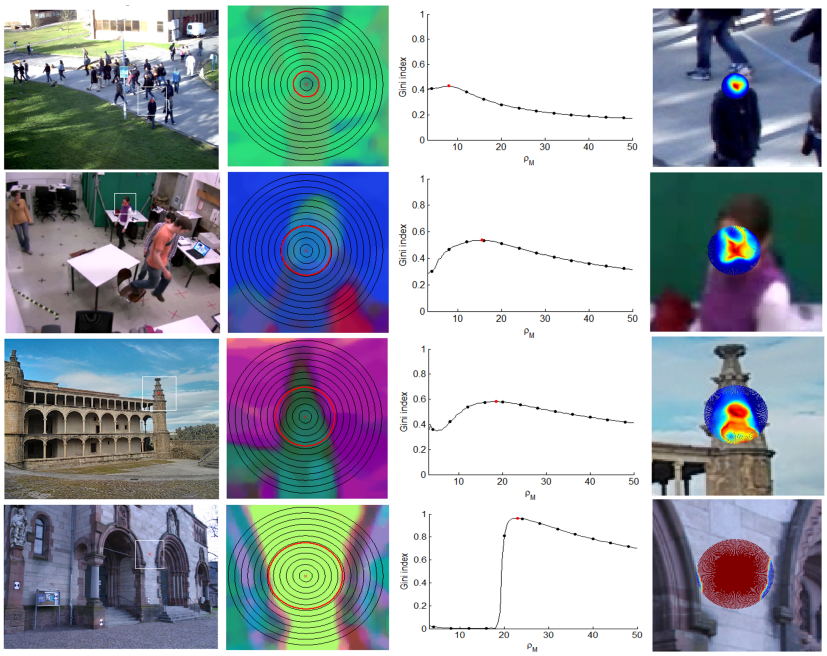


Fig. 9.4. Use of the Gini index to select the support size. For each row. First column: colour image with anchor point marked with a red cross and enlarged area enclosed by a white rectangle. Second column: first 3-components of image embeddings, (some of the) tested supports plotted as black circumferences and optimal support plotted in red. Third column: sparsity value measured by the Gini index for different support hypotheses: values for the shown black circumferences are indicated by black dots, the most-sparse configuration (the optimal) is indicated by a red dot. Fourth column: weights of the samples in the support, obtained by measuring their resemblance to the anchor point (the redder the higher, the bluer the lower). A $\gamma = 10$ has been used to improve visualization such that small differences can be better appreciated (observe the weights correlation with the embeddings in the second column). See extraction details in the text.

scale-space (e.g. the difference of Gaussians or the determinant of the Hessian matrix) is more prominent. The scale-space is built by successive convolutions of the image with a fixed-size Gaussian kernel. Points which are part of wide structures are prone to be detected at higher scales; therefore, these points will be described by severely smoothed images. In these smoothed images most of the detail is lost and the transitions between projected scene surfaces (object edges) might be blurred (in a pure scale-space) or displaced from its original position (in a pyramidal approximation of the scale-space as in SIFT). An alternative is to rely on a super-pixel segmentation process to define the description support—as in Micusik [2010]—. However, this usually conveys over-partitioned results, i.e. composed of homogeneous segments with insufficient representation capability. We require a description scheme which keeps image detail and accurately defines the support extent. Then, an alternative scheme for defining the support size

on the original image scale is needed. To this aim, we propose a MAP-scheme to automatically select the size of the support.

In this section we first present the generic shape of the support and define a generic characterization process with a set of simple features. Then, we review the resemble function proposed in Trulls et al. [2013] and present our strategy to automatically select the support size.

Point characterization

Description samples

The use of a polar grid to define an image point support is mainly motivated for its isotropic nature. Through polar sampling, the neighbours to describe the point are chosen at a uniform distance to the described point.

Given an image point $\boldsymbol{\psi} = (u, v, 1)^T$ let us define its support $\boldsymbol{S}_{\boldsymbol{\psi}}$ as the set of polar-sampled image points around it, including the point itself, such that the position of any point in the support, $\boldsymbol{\psi}_{j,k} = (u_{j,k}, v_{j,k}, 1)^T$ is given by:

$$u_{j,k} \coloneqq u + j\triangle\rho\, cos(k\triangle\theta)\ ,\quad v_j \coloneqq v + j\triangle\rho\, sin(k\triangle\theta) \tag{9.6}$$

In equation 9.6, $j\triangle\rho$ and $k\triangle\theta$, with $j \in [0, N_\rho - 1]$ and $k \in [0, N_\theta - 1]$, are the polar coordinates of the point $\boldsymbol{\psi}_{j,k}$ respect to $\boldsymbol{\psi}$. To obtain these coordinates, $\triangle\rho = \frac{\rho_M}{N_\rho - 1}$ and $\triangle\theta = \frac{2\pi}{N_\theta - 1}$ define a regular polar sampling of $N_\rho N_\theta$ samples on a circle of radius ρ_M around $\boldsymbol{\psi}$.

The support radius ρ_M fully defines the support size and, as aforementioned, it is automatically selected for each point according to the scene structure around the point. Hence, it can be different for different points in the image (see 9.4 for the selection strategy).

The number of support samples is obtained as a function of this radius as:

$$N_\rho \coloneqq N_\theta \coloneqq \left\lceil \frac{\sqrt{\pi(\rho_M)^2}}{D_S} \right\rceil \tag{9.7}$$

, where $\lceil \chi \rceil$ returns the closest integer bigger than χ and D_S is a parameter inversely proportional to the support sampling density—the higher: the lower the number of samples—. The sensitivity of the designed polar sampling scheme to D_S is inspected in section 9.6.

Description features

Our aim is to design a flexible characterization scheme, so that alternative features can be used in the future without altering the general idea of the present approach. Furthermore, we target to design a low-dimensional feature vector in order for the searching process (described in section 9.5) to be efficient. To this aim, we simply assume the features proposed by the authors

of NRDC HaCohen et al. [2011] but we change the comparison process there used (see section 9.5).

Let $f_{\mathbf{I}}(\boldsymbol{\psi}) : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ be a characterization function that returns, for every image point $\boldsymbol{\psi} = (u, v, 1)^T$ in the support, a 3-dimensional description vector, $\mathbf{Lab}(\boldsymbol{\psi})$, containing the luminance $\mathbf{I}_L(\boldsymbol{\psi})$ and chrominance $\mathbf{I}_a(\boldsymbol{\psi})$, $\mathbf{I}_b(\boldsymbol{\psi})$ values of image $\mathbf{I}$ at position (u, v). Additionally, let us define $g_{\mathbf{I}}(\boldsymbol{\psi}) : \mathbb{R}^2 \rightarrow \mathbb{R}^1$ be a characterization function which returns for every point in the support, the luminance gradient magnitude of the image $\mathbf{I}$ at position (u, v): $G(\boldsymbol{\psi})$.

The function $g_{\mathbf{I}}$ returns a description of the magnitude of the luminance variations in the support, whereas the function $f_{\mathbf{I}}$ is used to describe its colour properties. CIE-Lab colour space is used due to its ability to codify colour descriptions such that small perturbations of the colour can be almost equally perceived across the full representation range, i.e. CIE-Lab is almost perceptually uniform. The gradient phase is ignored as angular directions are not conserved by projective transformations.

Weighting-by-resemblance

The 3-dimensional structure of the scene is almost lost in its 2-dimensional projection: scene surfaces that are on different depth planes may be projected adjacent in the image. These projections create transitions in the image that are not related with real transitions between scene surfaces. These transitions are known as occlusion edges and their spatial position changes with the point-of-view. A description support for a point $\boldsymbol{\psi}$ may extend traversing the occlusion edges, including samples from another scene surface in the description of $\boldsymbol{\psi}$. We aim to detect the samples in the support which are projections of the surface on which $\boldsymbol{\Psi}$—the 3-dimensional retro-projection of $\boldsymbol{\psi}$—lies. To this aim, occlusion edges need to be somehow handled.

In Leordeanu et al. [2012], authors present a generic solution for edge detection by combining multiple cues. Among these cues, soft-segmentation appears as an efficient and robust mid-level representation to identify colour edge transitions that are unclear in the feature space—other features as texture may be also used w.l.o.g.—. In approximate terms their idea is to model the colours of any image patch as if they were generated from a distribution composed of a linear combination of a finite number of colour probability distributions. Under this idea, authors propose to extract a finite number of object/ground segmentations by estimating the latent colour distributions in the data. These segmentations are PCA processed and the top 8 components are selected per image point to represent a compressed version of the whole set of segmentations. Some examples of this soft-segmentation are depicted in the second column of Figure 9.4, where only the first 3 top components are shown combined in a single RGB-arranged image.

The so-extracted 8 components represent the colour distribution of the point respect to its neighbours. These components have been used for point description purposes in Trulls

et al. [2013]. Following the idea there presented, these components can be seen as an 8-dimensional embedding of each image point. Let us define $eb_{\mathbf{I}}(\boldsymbol{\psi}) : \mathbb{R}^2 \rightarrow \mathbb{R}^8$ as the embedding function that returns for the image point $\boldsymbol{\psi} = (u, v, 1)^T$ the 8-dimensional embedding obtained via the soft-segmentation process of image $\mathbf{I}$ at position (u, v). A composed function $e_{\mathbf{I},\boldsymbol{\psi}}(\boldsymbol{\psi}_{j,k}) : \mathbb{R}^2 \rightarrow \mathbb{R}^8 \rightarrow \mathbb{R}^1$ that quantifies the edge intensity between an image point $\boldsymbol{\psi}$ and one of its neighbours in the support $\boldsymbol{\psi}_{j,k}$ can be obtained by applying the $l^2 - norm$ between their embeddings:

$$e_{\mathbf{I},\boldsymbol{\psi}}(\boldsymbol{\psi}_{j,k}) \coloneqq |eb_{\mathbf{I}}(\boldsymbol{\psi}) - eb_{\mathbf{I}}(\boldsymbol{\psi}_{j,k})|_2 \tag{9.8}$$

Once the edge intensity is quantified, a conservative solution to include edge information in the description of a point $\boldsymbol{\psi}$ would search for points in the support with an embedding similar to $eb_{\mathbf{I}}(\boldsymbol{\psi})$ and discard the other points. This can be achieved by computing and applying a threshold to equation 9.8 for every neighbour. This would create a region around $\boldsymbol{\psi}$. The characterization of neighbours not assigned to this region would be leaved out of the description. This idea is similar to the one proposed in Navarro et al. [2014].

A fuzzy alternative is to weight the contribution of the neighbours according to the resemblance of their embedding to the embedding of the anchor point: $eb_{\mathbf{I}}(\boldsymbol{\psi})$. This is done in Trulls et al. [2013] by generating a vector of weights $\mathbf{w}_{\mathbf{I}}$ which value decreases exponentially with the edge intensity. For instance, for the point $\boldsymbol{\psi}_{j,k}$ in the support:

$$\mathbf{w}_{\mathbf{I}}(\boldsymbol{\psi}_{j,k}) \coloneqq exp(-\gamma e_{\mathbf{I},\boldsymbol{\psi}}(\boldsymbol{\psi}_{j,k})) \tag{9.9}$$

, where γ controls the decay of the exponential. These weights can be normalized $\dot{\mathbf{w}}_{\mathbf{I}}$ by simply dividing each weight in the vector by the summation of the weights in the support. We prefer this solution mainly due to its flexibility. However, note that for big values of γ, the effect of both solutions is almost equivalent. The sensitivity of this weighting scheme to γ is explored in section 9.6.

Support size selection

According to its definition in section 9.4, the circular support around $\boldsymbol{\psi}$ grows isotropically as ρ_M is increased. Intuitively, and assuming colour continuity on a plane's surface (texture patterns are not currently considered), the optimal value for ρ_M would be the one that maximises the number of samples in the support that are in the embedding of $\boldsymbol{\psi}$ and minimises the number of samples in other embeddings.

However, the inclusion of a few samples from adjacent embeddings may provide strong structural evidences of the scene surface, as they can be used to determine the extent of the surface and define the largest possible support. This structural information would be of main

interest in the appearance transformation of quasi-homogeneous supports (see a motivation for the use of the structural information in section 9.5). Nevertheless, the influence of the other embeddings in the description will be reduced by the weighting scheme.

Our solution aims to detect the value of ρ_M which derives supports on which most of it samples are placed in the same—or a similar—embedding of $\boldsymbol{\psi}$ but also contain a few samples placed in different embeddings. This configuration of the support would convey sparse distributions of the edge intensities in the support.

Intuitively, a sparse representation is one in which a small number of samples in the support contain a large proportion of the sum of edge intensities. Note that, by definition (see equation 9.8), edge intensities are always non-negative. In order to measure the sparsity of a set of values, in this case the set of edge intensities in the support, several methods have been compared in Hurley and Rickard [2009]. Among them, the Gini index satisfies a set of six criteria that ensures a desirable behaviour.

Being:

$\mathbf{e} = [e_{\mathbf{I},\boldsymbol{\psi}}(\boldsymbol{\psi}_{0,0}), .., e_{\mathbf{I},\boldsymbol{\psi}}(\boldsymbol{\psi}_{j,k}), ..., e_{\mathbf{I},\boldsymbol{\psi}}(\boldsymbol{\psi}_{N_\rho-1,N_\theta-1})]$, a vector of edge intensities, we first order the intensities from smallest to largest: $e(1) \leq e(2) \leq ..., \leq e(N_\rho N_\theta)$, where $(1), (2), ..., (N_\rho N_\theta)$ are the new indexes after the sorting process. The Gini index of the vector, $gi(\mathbf{e})$, is given by:

$$gi(\mathbf{e}) \coloneqq 1 - 2 \sum_{\kappa=1}^{N_\rho N_\theta} \frac{e(\kappa)}{|\mathbf{e}|_1} \left(\frac{N_\rho N_\theta - \kappa + \frac{1}{2}}{N_\rho N_\theta} \right) \tag{9.10}$$

We propose to test several radius hypotheses for each anchor point and compute the Gini index for the edge intensity vector resulting from each hypothesis. Finally, we select ρ_M as the radius value that maximises the Gini index. This process is illustrated for four different scenarios in Figure 9.4. Results in the Figure suggest that, for four remarkably different scenarios, the proposed solution is able to automatically locate the limits of the projected objects, hence defining appropriate description supports

The Gini index as defined in equation 9.10, is quite sensible to impulsive image noise. Let us imagine a situation in which a support with a particular radius is entirely composed of samples projected from a common surface. Furthermore, let be this surface homogeneous in colour. The edge intensities respect to the central sample should be close to zero for all the samples. In this scenario, even a low intense impulsive noise on one of the samples would convey highly sparse distributions. We can solve this problem by introducing a preliminary significance test on the edge intensities. In particular, radius hypotheses deriving in edge distributions not fulfilling equation 9.11 are rejected.

$$e(N_\rho N_\theta) - e(1) \geq 1 \tag{9.11}$$

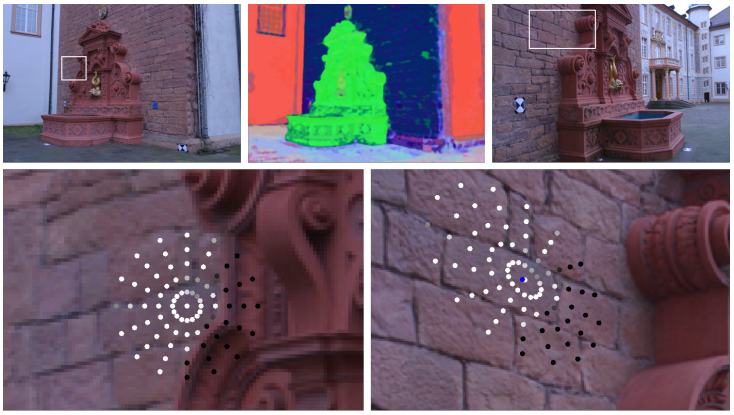


Fig. 9.5. Example of the proposed searching method. First row. Colour image $\mathbf{I}$ with the enlarged area enclosed by a white rectangle (top left). Segmentation-aware embeddings of the colour image (middle column). Another image $\mathbf{I}'$ of the same scene with the enlarged area enclosed by a white rectangle (top right). Second row left. Enlarged area in $\mathbf{I}$ with the anchor point and its description support around it. Resemblance of the support samples to the anchor point is here codified in grey-scale (the brighter the more resemble). Second row left. Enlarged area in $\mathbf{I}'$ with a geometrically transformed version of the support. The ground-truth matching for the anchor point is indicated by a blue dot. The support is distorted to adapt to its projective transformation. Influence of out-of-plane areas is diminished by segmentation-aware embedding. A small density sampling (big D_S value) has been used to ease visualization. Appearance of the support has been also transformed to allow the match (see Section 9.5). The small deviation to the ground-truth is due to the sampling of the normal and depth hypotheses spaces (see text for sampling details).

Relation with scale-space theory

Is is important to remark that it is not our aim to present an alternative to the scale-space theory, but to define a point support merely for descriptivism reasons, not for distinctiveness reasons, as it is the case of scale-space-like approaches. However, from a scale-space perspective, we can describe our approach as one that searches for saddle areas around the anchor point in a scale-space curvature in the image original scale.

9.5 Searching approach

Figure 9.5 exemplifies the proposed searching method. Our target is to spatially rearrange support samples under a projective transformation. In the figure, the support around the anchor point is adapted to the projective transformation of the wall (the surface on which it lies).

Under a projective transformation neither the size, nor the perimeter, nor the angles, nor the distance between samples is conserved. Only the cross-ratio (which involves two pairs of samples) is conserved. Note that the stability of the cross-ratio is automatically provided by the

arrangement of epipolar lines provided by the fundamental matrix.

Assuming that the support is mainly enclosed in a planar surface, the projective transformation that suffers can be modelled by an 8-parameter homography (see equation 9.4). If the scene is fully calibrated, the whole family of possible homographies is function of just the three scalar parameters that fully define a scene plane:

- The distance of the scene plane to the focal plane of the camera $\mathbf{C}$ that captures the image $\mathbf{I}$, related toλ_Π as explained in section 9.3.
- The azimuth component of the scene plane orientation: n_θ.
- The elevation component of the scene plane orientation: n_φ.

The normal vector of the scene plane that defines the orientation has been here expressed in spherical coordinates $\boldsymbol{n}_{\boldsymbol{\Pi}} = (n_\theta, n_\varphi, 1)$. Note that the radial distance equals 1 as the normal was defined to be of norm unitary (see section 9.3).

The problem of finding the correct homography is two-fold: finding the vector in the unit sphere that better describes the plane orientation (n_θ, n_φ); and finding the distance of such plane to the camera (λ_Π if its fronto-parallel). If both problems are solved the result is also two-fold: an homography that maps points in the plane's projections in $\mathbf{C}$ and $\mathbf{C}$' , and relevant information about the 3-dimensional scene.

Projective transformation of the support

The set of possible plane-induced homographies is defined by the corresponding set of possible vectors of parameters, so that every hypothesis $\mathbf{h} = \{\lambda_\Pi^{(h)}, n_\theta^{(h)}, n_\varphi^{(h)}\}$ defines a plane-induced homography $\mathbf{H}(\mathbf{h}) = \mathbf{H}(\lambda_\Pi^{(h)}, \boldsymbol{n}_{\boldsymbol{\Pi}}{}^{(\mathrm{h})})$. For a given hypothesis $\mathbf{h}$, $\boldsymbol{n}_{\boldsymbol{\Pi}}{}^{(\mathrm{h})}$ should be first converted to Cartesian coordinates from its polar form. The corresponding homography is then used to project each description sample $\boldsymbol{\psi}_{j,k}$ in the anchor support $S_{\boldsymbol{\psi}}$ in image $\mathbf{I}$ (by applying equation 9.5), which results in a projected support $S'_{\boldsymbol{\psi}}(\mathbf{h})$ in $\mathbf{I}'$. Therefore, $S'_{\boldsymbol{\psi}}(\mathbf{h})$ is a $N_\rho N_\theta$-set of projected samples, $\boldsymbol{\psi}'_{j,k}(\mathbf{h})$, that index image coordinates on $\mathbf{I}'$ under the hypothesis $\mathbf{h}$.

The projection of the points in the projected support, $S'_{\boldsymbol{\psi}}(\mathbf{h})$, is a function of both the normal and depth hypothesis. In essence, for a depth hypothesis the anchor point is projected onto the $\mathbf{I}'$ image on a particular position. Next, for the associated normal hypothesis, the projected position of this point is kept unaltered whereas the other points in the support (the description samples) are projected around it. These other projected points define new positions on which descriptions will be extracted. If we observe the position of one of these projected points for different normal hypothesis, we would see the point orbiting around the projection of the anchor point.

The interpretation of this scheme may be better understood by using the SIFT description as a reference. The SIFT descriptor is aligned with the dominant orientation(s) in the SIFT support. This technique achieves robustness against potential potential 2-dimensional rotations of the support. We propose to extend this by attending also to 3-dimensional rotations. As we aim to operate without any knowledge of the 3-dimensional scene, instead of aligning the description, we consider 3-dimensional rotations in the descriptor matching process.

The aim of the process is to find the optimal hypothesis $\mathbf{h}^*$ for which $\mathbf{H}(\mathbf{h}^*)$ projects the centre of the support (i.e. the anchor point $\boldsymbol{\psi}$ on image $\mathbf{I}$) on its *ground-truth* position $\boldsymbol{\psi}'(\mathbf{h}^*)$ on image $\mathbf{I}'$ and properly transform its support. Once the correct correspondence is established, the orientation of the plane $\boldsymbol{n}_{\boldsymbol{\Pi}}$ on which their 3-dimensional projection $\boldsymbol{\Psi}(\mathbf{h}^*)$ lies, and the distance λ of $\boldsymbol{\Psi}(\mathbf{h}^*)$ to $\mathbf{C}$, computed from $\lambda_{\Pi}^{(\mathbf{h}^*)}$, are also automatically obtained.

This is the same process followed by classical plane sweeping strategies. However, these strategies only sweep on the depth parameter and on a subset of the possible plane orientations, as a *Manhattan* (or *Atlanta*) world is assumed. We instead target a generic scenario, so that the set of hypothesis is just limited by geometrical restrictions: scene calibration and images extent provide strong geometrical constraints to reduce the number of hypotheses. These constraints are, to our knowledge, first used together in the proposed approach.

Limiting the set of hypothesis

Reducing the number of hypotheses would obviously reduce the computational cost of the searching stage, as the size of the family of homographies to test would decrease. Furthermore, the geometrical constraints, if successfully applied, also reduce the likelihood of finding false positive correspondences (*distractors*).

The principle of our constraining process is: *do not consider hypothesis that do not fit to the capture conditions*. Two different strategies are proposed to constraining the depth and normal ranges:

- Although we cannot determine *a priori* which is the closest (or furthest) surface to the $\mathbf{C}$ camera and the distance at which such surface *is* placed, we can instead obtain the minimum (and maximum) distance at which a scene surface *has to be* placed to be observable by the other camera $\mathbf{C}'$.

- As the cameras positions are assumed to be fixed, we can impose constrains on the surfaces orientations. An orientation is considered if the associated surface shows the same side to both cameras.

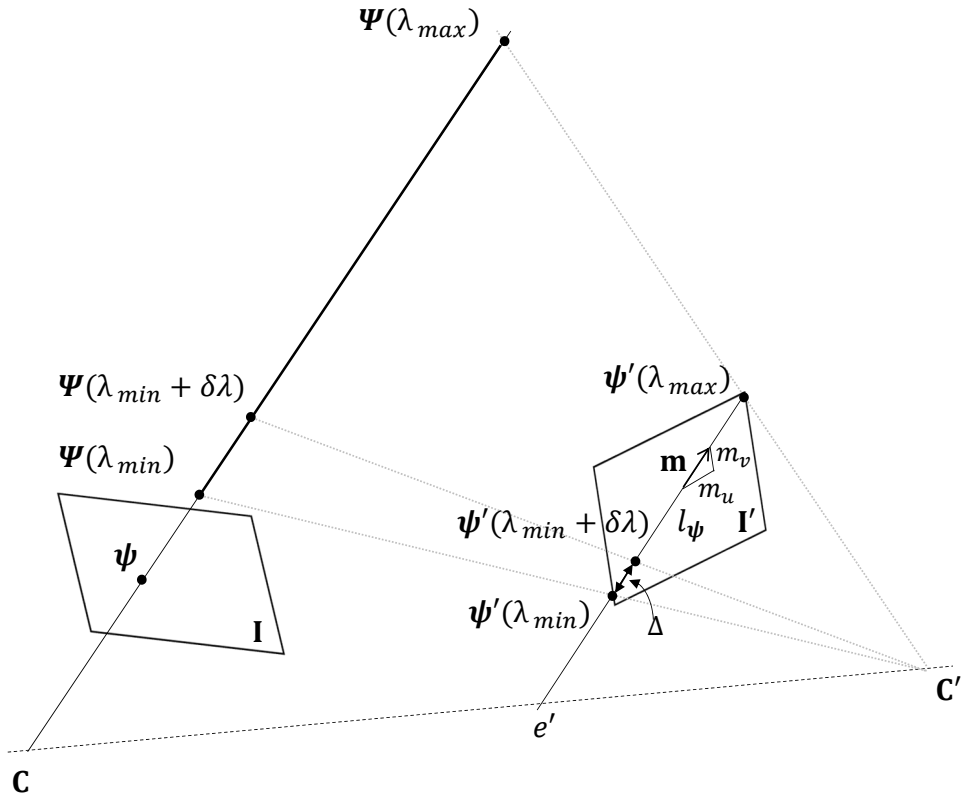


Fig. 9.6. Constraining and sampling the depth range. For a point $\boldsymbol{\psi}$ on image $\mathbf{I}$ the range of depths $\lambda_{\Pi}^{(h)} \in [\lambda_{min}, \lambda_{max}]$ for which its back-projections are observable on $\mathbf{I}'$ is constrained by the extent of $\mathbf{I}'$. Additionally, a non-uniform sampling of the depth range for $\boldsymbol{\psi}$ can be derived from a uniform sampling of its epipolar line $\boldsymbol{l}_{\boldsymbol{\psi}}$; given the line direction vector, $\mathbf{m}$, and the desired spatial sampling, Δ, on the epipolar line. See text for details.

Constraining the depth range

The observable depth range for a point $\boldsymbol{\psi}$ can be constrained by its epipolar line $\boldsymbol{l}_{\boldsymbol{\psi}}$. Specifically, by the cuts of $\boldsymbol{l}_{\boldsymbol{\psi}}$ with the spatial limits of image $\mathbf{I}'$. Matching of $\mathbf{I}$ image samples which projections fall out of the extent of $\mathbf{I}'$ cannot be determined by comparison (see Figure 9.6).

Let us assume that we start from a point $\boldsymbol{\psi}'(\lambda_{\Pi}) = (u'(\lambda_{\Pi}), v'(\lambda_{\Pi}), 1)^T$ on $\mathbf{I}'$, which is the projection of a point $\boldsymbol{\psi}$ on $\mathbf{I}$ under a depth hypothesis: λ_{Π}. $\boldsymbol{\Psi}(\lambda_{\Pi})$, the 3-dimensional point obtained by back-projecting $\boldsymbol{\psi}$ at such depth hypothesis can be obtained by computing the intersection between the optical rays emanating from $\mathbf{C}$ and $\mathbf{C}'$. Ideally, these rays would intersect exactly at the same 3-dimensional point. However, as the camera parameters are most of the times only known approximately, these rays may not intersect. So instead, our aim is to seek for the 3-dimensional point that has a minimum distance from both rays. This point would be $\boldsymbol{\Psi}(\lambda_{\Pi})$ and can be obtained by triangulation, for what we use the Marquardt algorithm (as

in Aanaes et al. [2012]). Given $\boldsymbol{\Psi}(\lambda_\Pi)$, the depth hypothesis at which it is back-projected can be estimated by isolating λ_Π in equation 9.2.

This triangulation process might be used to extract λ_{min} and λ_{max} for a given point $\boldsymbol{\psi}$ with $\boldsymbol{\psi}'(\lambda_{min})$ and $\boldsymbol{\psi}'(\lambda_{max})$ being the intersections of the point's epipolar line $\boldsymbol{l}_{\boldsymbol{\psi}}$ with the spatial limits of $\mathbf{I}'$. Then, the range for the depth parameter to test is constrained on a closed interval: $\lambda_\Pi^{(h)} \in [\lambda_{min}, \lambda_{max}]$.

Sampling the depth range

Small displacements along the epipolar line might correspond to large displacements along the optical ray connecting $\boldsymbol{\Psi}(\lambda_\Pi)$ and $\mathbf{C}$. As captured images are discrete signals, the epipolar line is not a continuous line. Instead, it is somehow uniformly sampled according to the image resolution and to the line slope. Hence, when analysing the possible values of the $\lambda_\Pi^{(h)}$ parameter, we may consider only those corresponding to uniformly sampled points on the epipolar line. This leads to a non-uniform sampling of the optical ray. Nevertheless, the uniform sampling of the epipolar line can be also done w.l.o.g. on non discrete steps, achieving what is known as sub-pixel precision. We propose here a configurable scheme to achieve a non-uniform sampling of the optical ray (sampling of the depth range) by defining a configurable uniform sampling of the epipolar line, with, r, the sampling parameter.

Let us define $\mathbf{m} = (m_u, m_v)$ as the direction vector of the epipolar line $\boldsymbol{l}_{\boldsymbol{\psi}}$. Given a projection of the point $\boldsymbol{\psi}$: $\boldsymbol{\psi}'_1 = (x', y', 1)^T$ and another projection also on $\boldsymbol{l}_{\boldsymbol{\psi}}$ at a distance Δ from this point: $\boldsymbol{\psi}'_2 = (u' + \Delta m_u, v' + \Delta m_v, 1)^T$ the distance (or depth increment) $\delta\lambda$ between their corresponding scene points can be obtained by triangulating from $\boldsymbol{\psi}$ and $\boldsymbol{\psi}'_1$ to obtain $\lambda_\Pi^{(h)}$ and from $\boldsymbol{\psi}$ and $\boldsymbol{\psi}'_2$ to obtain $\lambda_\Pi^{(h)} + \delta\lambda$ (see Figure 9.6).

Therefore, the idea is to generate a uniform sampling of the epipolar line starting from the first cut with $\mathbf{I}'$ and ending in the other cut:

$\boldsymbol{\psi}'(\lambda_{min}), ..., \boldsymbol{\psi}'(\lambda_\Pi^{(h)}), ...\boldsymbol{\psi}'(\lambda_{max})$, such that for every two consecutive points:

$$\left|\boldsymbol{\psi}'(\lambda_\Pi^{(h+1)}) - \boldsymbol{\psi}'(\lambda_\Pi^{(h)})\right|_2 = \Delta \tag{9.12}$$

By triangulation, this process results in: $(\lambda_{min}, ..., \lambda_\Pi^{(h)}, ...\lambda_{max})$, an ordered set of depth possible values for each anchor point $\boldsymbol{\psi}$. The process is illustrated in Figure 9.6 for $\lambda_\Pi^{(h)} = \lambda_{min}$. A similar depth sampling process is proposed in Tola [2010].

Constraining the range of orientations

For a given point and a possible depth value only a subset of the plane orientations are observable in both views. This idea can be easily exemplified by a thin surface in the scene, e.g. a traffic sign. If in a wide-baseline scenario $\mathbf{C}$ and $\mathbf{C}'$ are placed in a way that they observe different faces of

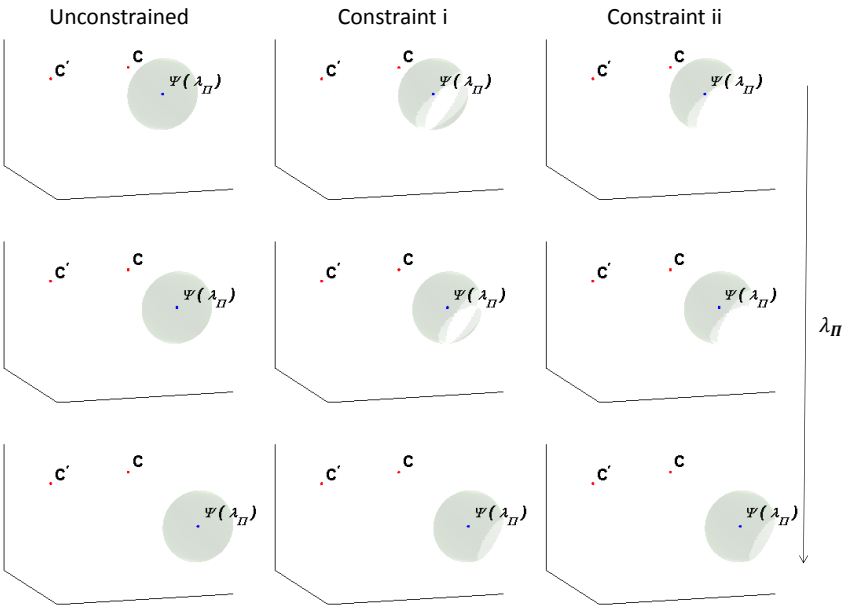


Fig. 9.7. Constraining the range of scene planes orientations. The relative position between the cameras $\mathbf{C}$ and $\mathbf{C}'$ strongly constrains the observable plane orientations which are represented via the unitary sphere. In an unconstrained scenario the whole space of orientations is considered. As both cameras should observe the same side of the plane (constraint i) some of the orientations are discarded. If orientations are reduced to those that describe planes in front of both cameras (constraint ii) additional normals are discarded. However, orientation restriction decrease as the depth of the plane (λ_Π) as can be observed comparing a row with the row below. See text for details.

the surface and as the faces might be different, there will be no image evidences to match points between the surface projections. There are solutions to aggregate inconsistent correspondences in these areas (Ummenhofer and Brox [2013]), but a preliminary correspondence map is first required.

The normal constraining scheme that we propose builds on this idea to limit the considered plane orientations. Let us divide the scheme into two constraints to ease the understanding of the idea. The first constraint (constraint i in the Figure 9.7) imposes the requirement that both cameras must see the same face of the plane. Let us express the camera centres as 3-dimensional points: $\mathbf{c} = (c_U, c_V, c_W, 1)^T$ and $\mathbf{c}' = (c'_U, c'_V, c'_W, 1)^T$ and the hypothesised plane by its general equation: $\mathbf{\Pi}^{(h)} = (n_U^{(h)}, n_V^{(h)}, n_W^{(h)}, \lambda_\Pi^{(h)})^T$.

Both cameras capture the same face of the plane if:

$$sign(\mathbf{\Pi}^{(h)\mathrm{T}} \times \mathbf{c}) = sign(\mathbf{\Pi}^{(h)\mathrm{T}} \times \mathbf{c}') \tag{9.13}$$

, where $sign(\chi)$ returns -1 if $\chi < 0$ and 1 otherwise.

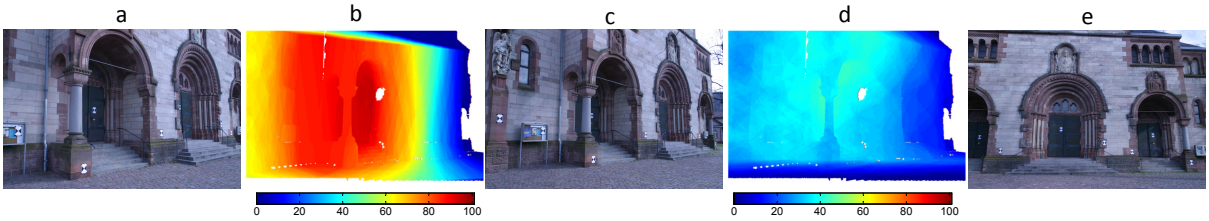


Fig. 9.8. Effect of the constraint of the scene planes orientations. Percentage of normals analysed (the redder the higher, the bluer the lower) when searching points of image c on images a (b) and e (d). Note how the number of analysed normals decreases as the anchor points are further to the image centre (both in b and d). Furthermore, note how the percentage of analysed normals drastically decreases with the cameras separation (compare b and d). Normal constraining has been performed at ground truth depth. Image points without ground-truth depth information (white areas) are leaved out of the analysis. Images extracted from the data-set proposed in Strecha et al. [2008].

However, this constraint does not account for plane orientations that point away from the cameras. The projections of the surface on the images only represent the visible face of the surface, i.e. the back face of the surface is hidden to the camera by the surface itself. To this aim, only surfaces which normals point towards the cameras configure hypothesised planes which are observable on both views. This represent our second constraint (constraint ii in the Figure 9.7) and is simply obtained by imposing:

$$sign(\mathbf{\Pi}^{(h)\mathrm{T}} \times \mathbf{c}) = sign(\mathbf{\Pi}^{(h)\mathrm{T}} \times \mathbf{c}') = 1 \tag{9.14}$$

The normal hypothesis is observable (and thus plausible) if and only if, equations 9.13 and 9.14 hold.

The number of possible orientations is proportional to: the relative position between the cameras, the further the lower; the position of the image point, the further from the centre the lower (both effects are shown in Figure 9.8); and the distance to the camera centre of the 3-dimensional projection of the image point, the higher the lower the number of discarded normals (illustrated in Figure 9.7).

Sampling the orientation range

In order to also sample the orientation, we introduce an additional parameter $D_{\boldsymbol{n}}$, that defines the number of solid angles in which the unit sphere is uniformly divided. Considered orientations, for each point $\boldsymbol{\psi}$ and every considered depth value λ_{Π}, are obtained by sampling the azimuth and elevation ranges $[0, 2\pi]$ in angular steps $\lceil \sqrt{D_{\boldsymbol{n}}} \rceil$. The sensitivity of the designed orientation sampling scheme to $D_{\boldsymbol{n}}$ is inspected in section 9.6.

Handling description integrity

Each hypothesis $\mathbf{h} = \{\lambda_{\Pi}^{(h)}, n_{\theta}^{(h)}, n_{\varphi}^{(h)}\}$ results in a candidate projection of the support $S_{\boldsymbol{\psi}}$ on image $\mathbf{I}'$: $S'_{\boldsymbol{\psi}}(\mathbf{h})$. We can use the description functions defined in section 9.4 to extract features on each sample $\boldsymbol{\psi'}_{j,k}(\mathbf{h})$. In particular, we extract the CIE-Lab colour description $\mathbf{Lab}(\boldsymbol{\psi'}_{j,k}(\mathbf{h}))$ through the function $f_{\mathbf{I}'}(\boldsymbol{\psi'}_{j,k}(\mathbf{h}))$ and the gradient magnitude $G(\boldsymbol{\psi'}_{j,k}(\mathbf{h}))$ through the function $g_{\mathbf{I}'}(\boldsymbol{\psi'}_{j,k}(\mathbf{h}))$ for every point in the candidate projected support $S'_{\boldsymbol{\psi}}(\mathbf{h})$ on image $\mathbf{I}'$. However, the appearance of the support may be altered due to wide-baseline effects even if $\mathbf{h}$ represents a reliable geometrical transformation of the support (see second row of Figure 9.9).

In order to cope with these appearance changes, we follow the idea proposed in HaCohen et al. [2011] and design a local appearance transformation model. However, we enhance their model by including resemblance information (weights) in the calculation of the transformation model. The goal of this process is to avoid the influence of occluding and background samples in the anchor support. The result of this process is the generation of transformed descriptions $\mathcal{T}_f(\mathbf{Lab}(\boldsymbol{\psi'}_{j,k}(\mathbf{h})))$ and $\mathcal{T}_g(G(\boldsymbol{\psi'}_{j,k}(\mathbf{h})))$ for every point in the projected support. To define the surface-aware appearance transformation model, we compute the optimal gain (ν) an bias (η) between the description features $\{L, a, b, G\}$ in the anchor support $S_{\boldsymbol{\psi}}$ and those in the candidate support, $S'_{\boldsymbol{\psi}}(\mathbf{h})$. For instance, for the gradient magnitude:

$$\nu_G(\mathbf{h}) \coloneqq \frac{\sigma_G(S_{\boldsymbol{\psi}})}{\sigma_G(S'_{\boldsymbol{\psi}}(\mathbf{h}))} \tag{9.15}$$

and:

$$\eta_G(\mathbf{h}) \coloneqq \mu_G(S_{\boldsymbol{\psi}}) - \nu_G(\mathbf{h})\mu_G(S'_{\boldsymbol{\psi}}(\mathbf{h})) \tag{9.16}$$

, where μ_G and σ_G are the weighted mean and weighted standard deviation for the gradient feature in the supports obtained using the normalized weights of the anchor support $\dot{\mathbf{w}}_{\mathbf{I}}$. For the projected support, these are obtained by:

$$\mu_G(S'_{\boldsymbol{\psi}}(\mathbf{h})) = \sum_{j,k} \dot{\mathbf{w}}_{\mathbf{I}}(\boldsymbol{\psi}_{j,k}) G(\boldsymbol{\psi'}_{j,k}(\mathbf{h})) \tag{9.17}$$

$$\sigma_G(S'_{\boldsymbol{\psi}}(\mathbf{h})) = \frac{\sum_{j,k} \dot{\mathbf{w}}_{\mathbf{I}}(\boldsymbol{\psi}_{j,k})(G(\boldsymbol{\psi'}_{j,k}(\mathbf{h})) - \mu_G(S'_{\boldsymbol{\psi}}(\mathbf{h})))}{\frac{(N_{w_{\mathbf{I}}}-1)}{N_{w_{\mathbf{I}}}}} \tag{9.18}$$

, where $N_{w_{\mathbf{I}}}$represents the number of non-zero weights in $\dot{\mathbf{w}}_{\mathbf{I}}$. Note that we use the weights of the anchor support to reduce the influence of transformed support samples which are occluded in this anchor support, not in the projected candidate support. The so-obtained gain and bias are used to define the appearance transformation models:

$$\mathcal{T}_g(G(\boldsymbol{\psi'}_{j,k}(\mathbf{h}))) = \nu_{\mathbf{G}}(\mathbf{h}) G(\boldsymbol{\psi'}_{j,k}(\mathbf{h})) + \eta_G(\mathbf{h}) \tag{9.19}$$

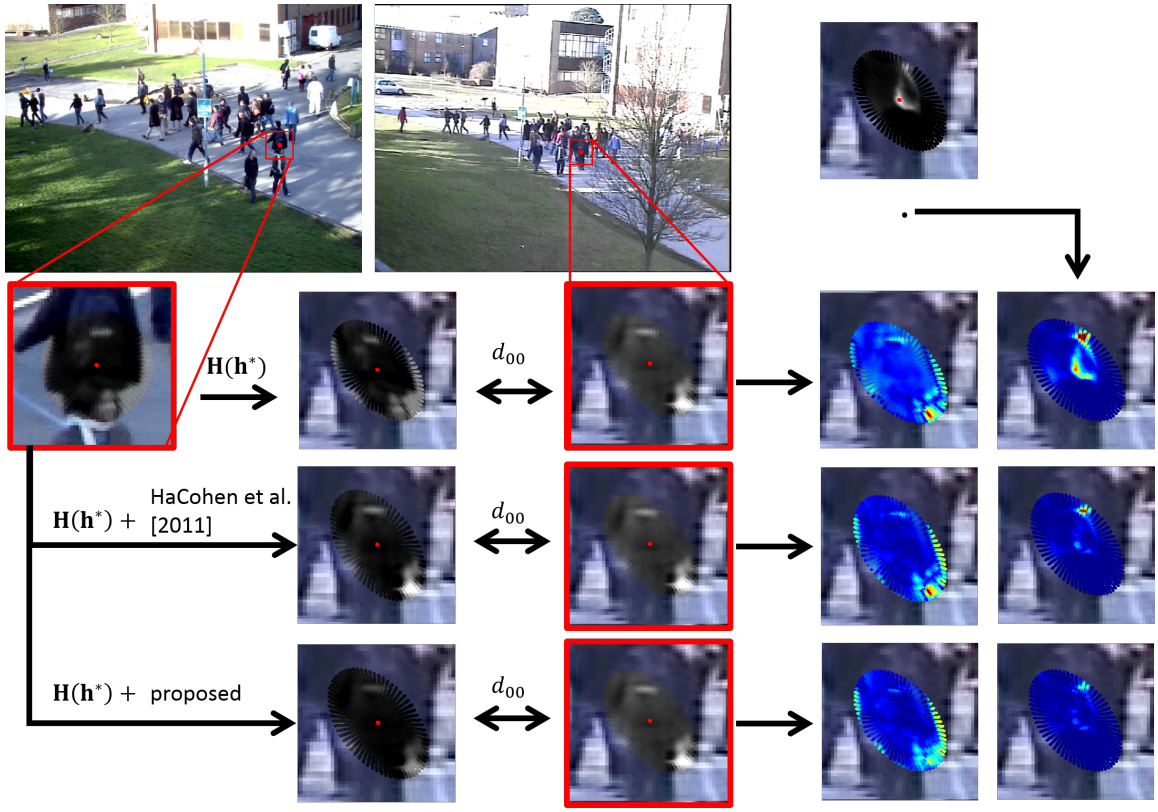


Fig. 9.9. Surface-aware appearance transformation. Only luminance information is shown (in grey-scale over-imposed on the colour image). Distances are colour-codified to ease visualisation (the darker the lower). First row (left to right): input image $\mathbf{I}$ and an anchor point, $\boldsymbol{\psi}$ (red dot); image $\mathbf{I}'$ and the corresponding point $\boldsymbol{\psi}'(\mathbf{h}^*)$; weights in the source support arranged according to the projection defined by $\mathbf{H}(\mathbf{h}^*)$. Second row (left to right): close-up of anchor area and anchor support $S_{\boldsymbol{\psi}}$; transformed support $S'_{\boldsymbol{\psi}}(\mathbf{h})$; close-up of corresponding area on $\mathbf{I}'$; distance without appearance transformation; distance after weighting with $\gamma = 50$. Third row, same as second row but applying the appearance transformation scheme in HaCohen et al. [2011] after geometric transformation. Fourth row, same as second and third row but applying the proposed appearance transformation method. See text for discussion.

$$\mathcal{T}_f(\mathbf{Lab}(\boldsymbol{\psi'}_{j,k}(\mathbf{h}))) = \mathbf{V}_f(\mathbf{h}) \times \mathbf{Lab}(\boldsymbol{\psi'}_{j,k}(\mathbf{h})) + \boldsymbol{\eta}_f(\mathbf{h}) \tag{9.20}$$

, where $\mathbf{V}_f(\mathbf{h}) \coloneqq diag(\nu_L(\mathbf{h}), \nu_a(\mathbf{h}), \nu_b(\mathbf{h}))$ is the diagonal matrix of the CIE-Lab colour gains and $\boldsymbol{\eta}_f(\mathbf{h}) \coloneqq (\eta_L(\mathbf{h}), \eta_a(\mathbf{h}), \eta_b(\mathbf{h}))^T$ its associated bias vector. For the shake of simplicity we have assumed linear independence among the features.

Figure 9.9 illustrates the benefits of the surface-aware appearance transformation model when comparing it with a version of the searching approach without appearance transformation and a version of the searching approach using the solution described in HaCohen et al. [2011]. For visualization purposes, only transformed luminance information is shown, albeit chroma

and gradient information is also transformed. The colour distance to the anchor support is measured in terms of the d_{00} distance, which was used in chapter 4 and is again described in section 9.5. In Figure 9.9, note how the woman's face in the bottom of the support is partially eliminated by HaCohen et al. [2011] and almost completely eliminated by the proposed method. Additionally, see how the position of the white trademark on the bag is altered by both appearance transformation schemes. Obtained distances before and after weighting are lower through the proposed appearance transformation scheme, because of the lower influence in the transformation parameters of the lit area in the bottom part of the projected candidate support.

Transform over-fitting

In HaCohen et al. [2011] the bias and gain ranges defining the transformation were constrained to reduce the transformation capability of the scheme. The constraints were applied in order to avoid the over-fitting of the appearance transformation model. Differently to HaCohen et al. [2011] we do not impose any constraint on the gain an bias. Whereas this choice enhances the flexibility of the transform, it may be counterproductive when adapting a feature-homogeneous support with close to 0 feature deviations. Note that, in this case, the numerator of equation 9.15 tends to 0; hence, the gain tends also to 0 and the bias (equation 9.16) towards the mean of the feature in the support.

In this situation, the appearance transformation model will assign to all the samples in the projected support the mean of the feature in the anchor support. Therefore, for this feature, the distance between the supports would be small, with independence of the features in the transformed support.

In order to face this situation, we have introduced a few boundary samples in the support selection technique (as explained in section 9.4). Through this scheme, we ensure that description supports are not entirely homogeneous.

Matching the anchor support and the candidate supports

In this section we describe the metrics used to measure the goodness of of each projection hypothesis, represented by a projected candidate support. In particular, we derive two metrics to compare the anchor support with each candidate support: colour d_{color}, gradient d_{grad}. These are based on the appearance features described in section 9.4 and on the weighting scheme described in section 9.4.

The colour metric d_{colour} relies in the computation of the CIEDE00 (d_{00}) distance between the CIE-Lab descriptions of the samples in the anchor support and their candidate projections after feature transformation. This metric is preferred to the $l^2 - norm$ (used in HaCohen et al. [2011]) due to its superior behaviour in measuring changes for small colour differences (Habekost [2013]).

In order to shun the influence of occlusions and background samples in the calculation of the metric, the contribution of each support sample is weighted by its resemblance to the anchor point. Under this scheme, d_{colour} can be defined:

$$d_{colour}(\mathbf{h}) \coloneqq \sum_{j,k} \dot{\mathbf{w}}_{\mathbf{I}}(\boldsymbol{\psi}_{j,k})\, d_{00}(\mathbf{Lab}(\boldsymbol{\psi}_{j,k}), \mathcal{T}_f(\mathbf{Lab}(\boldsymbol{\psi'}_{j,k}(\mathbf{h})))) \tag{9.21}$$

Similarly, the gradient metric d_{grad} is obtained as a weighted version of the l^1-norm between the gradient magnitude descriptions of the support and the hypothesized projected support after appearance transformation:

$$d_{grad}(\mathbf{h}) \coloneqq \sum_{j,k} \dot{\mathbf{w}}_{\mathbf{I}}(\boldsymbol{\psi}_{j,k}) \left| G(\boldsymbol{\psi}_{j,k}), \mathcal{T}_g(G(\boldsymbol{\psi'}_{j,k}(\mathbf{h}))) \right|_1 \tag{9.22}$$

Obtaining the optimal hypothesis

The optimal hypothesis $\mathbf{h}^*$ is obtained by minimizing any of the two defined distances; for instance for the colour distance:

$$\mathbf{h}^* \coloneqq \underset{\mathbf{h}}{argmin}(d_{colour}(\mathbf{h})) \tag{9.23}$$

Hence, different hypothesis may be obtained for each metric. The operation of the two distances is compared in section 9.6.

9.6 Experimental results

In this section we evaluate the proposed method for the task of point matching. First, we present the data-set and measure its complexity by exploring the performance of SoA methods. Then, we describe the configuration of the methods used for comparison. The configuration parameters of our method are selected by a sensitivity analysis. The so-configured method is compared with state-of-the-art methods in quantitative and qualitative terms.

Data set description

The designed method requires a calibrated scenario to operate. Although there are plenty of calibrated stereo data-sets for research evaluation purposes, the availability of data-sets presenting wide-baseline scenarios is scarcer. Our goal is to evaluate the designed approach in scenarios

of varied nature in terms of: size and appearance of the captured objects; background texture; inter-camera separation and illumination conditions. To this aim, we make use of four different data-sets in which calibration information is (at least roughly) available. Selected scenes are shown in Figure 9.1.

The first two scenes, *fountain* and *herzjesu*, are part of the reconstruction data-set proposed in Strecha et al. [2008]. This data-set was designed as a benchmarking solution to evaluate image-based rendering methods. Each scene in the data-set is composed of a set of images captured from small-separated cameras positions. The scenes are fully and accurately calibrated. For the reported experiments, we have selected the two most-separated views which content partially overlap. Even though these scenes do not contain natural objects nor a huge number of orientation surfaces, their use is motivated by the presence of strong occlusions and by their associated depth maps. These maps are used to evaluate the goodness of our method in estimating the distance to the camera of the anchor points. Furthermore, in these two scenes, illumination conditions have been manually altered to evaluate the benefits of the appearance transformation module.

The next three scenes, *greens*, *fabric* and *wood*, are selected from the point-matching data-set proposed in Aanaes et al. [2012]. This data-set aims to evaluate existing point-of-interest detectors in a realistically challenging data-set. We have chosen three scenes which contain natural surfaces, repetitive textures and strong occlusions and illumination changes. Captured scale is high, and images are strongly detailed. Again, the two most-separated views available were selected for each scene. The three scenes are fully and accurately calibrated. However, depth information is not available, and 3-dimensional surface points are only partially supplied.

Finally, the last scenes, *indoors* and *outdoors*, were obtained from the surveillance data-sets described in Possegger et al. [2013] and Ferryman et al. [2009] respectively. These scenes present some challenging factors: image quality is worse than in the other scenes; objects are captured at medium to small scales and their details are lost in the process. In the outdoors scene, the two cameras analysed capture the scene with strongly different gain parameter, resulting in severe illumination-changed views. Additionally, these scenarios are only roughly calibrated, and epipolar constrains are not fulfilled for some of the image points. Depth information is not available for these data-sets.

Data-set complexity

In order to objectively measure the complexity of the data-set, we evaluate the performance of autonomous SoA methods (those that do not require calibration) on each scene. On one hand, for sparse methods (SIFT, LIOP and ASIFT) we measure the number of correspondences found (M) and the number of them that are consistent with the epipolar constraint (E). A match is considered consistent with the epipolar geometry if the euclidean distances of the involved points

Method / Scene		fountain	herzjesu	greens	fabric	wood	indoors	outdoors
SIFT	M	123	270	469	434	815	689	136
	E	11	84	118	19	389	9	17
	%	**8.9**	**31.1**	**25.2**	**4.4**	**47.7**	**1.3**	**12.5**
LIOP	M	88	125	240	541	416	674	139
	E	4	11	7	1	27	6	2
	%	**4.5**	**8.8**	**2.9**	**0.2**	**6.5**	**0.9**	**1.4**
ASIFT	M	0	79	61	22	76	0	16
	E	0	59	15	5	37	0	6
	%	**-**	**74.7**	**24.6**	**22.7**	**48.7**	**-**	**37.5**

Table 9.3: Data-set complexity according to state-of-the-art performance in terms of number of detected correspondences (M), number of correspondences consistent with epipolar constraints amongst the detected (E) and representation percentage (%) of E respect to M (basically, 100E/M). See text for details and discussion.

to their epipolar lines is smaller than 10 pixels. Results for these evaluations are included in Table 9.3.On the other hand, for dense methods (SSID, NRDC) we include their obtained image wrappings (for SSID, these are obtained through the SIFT-flow algorithm Liu et al. [2011]) in Figure 9.10.

Source codes have been obtained from Vedaldi and Fulkerson [2008] (SIFT and LIOP), and associated software to Yu and Morel [2009] (ASIFT), Trulls et al. [2013] (SSID), and HaCohen et al. [2011] (NRDC). The Harris-Laplace detector with affine adaptation and orientation detection has been used for both SIFT and LIOP descriptors. Images from some scenes (*greens*, *fabric* and *woods*) have been resized for ASIFT, SSID and NRDC due to codes limitations in processing big images. In particular, matching coordinates obtained by ASIFT are re-scaled to the original image resolution. Default configuration parameters have been used for all the evaluated methods.

In the light of Table 9.3 and Figure 9.10 we can conclude that none of the evaluated methods is able to find accurate and numerous correspondences for all the scenes. In fact, their performance in some scenes establish a wide space for improvement.On one hand, under the SIFT description a big number of correspondences are found, albeit only a moderate proportion of them are consistent with the epipolar geometry (which is a necessary, but not sufficient, condition for correctness). The use of LIOP on the same points (both descriptors share the detector) results in substantially worse numbers. This seems to indicate that LIOP, under the default configuration, is unable to face the proposed scenarios. ASIFT provides more stable results: whereas the

method produces a smaller number of correspondences than SIFT (as smaller as zero for two of the scenes), a higher proportion of them are coherent with the epipolar geometry. NRDC provides accurate but very sparse correspondences. Finally, the results from the combination of SSID and SIFT-flow range from decent (*wood*) to useless (*indoors*) reconstructions.

According to the scene complexity, results indicate that *herzjesu*, *greens* and *wood* appear to be the less complex scenes. Even though, obtained results are not accurate enough to be considered significant. On the other hand, *fountain*, *indoors*, *fabric* and, in a lower extend, *outdoors*, constitute scenes that are too complex to be faced by the explored autonomous SoA solutions.

Compared approaches

As discussed in previous section, none of the SoA evaluated algorithms is able to operate effectively on the proposed data-set. Nevertheless, this operation performance was expected. The data-set contains projective-related challenges that are out-of-the-scope of evaluated methods. In fact, this situation motivated the design of the proposed method.

In the proposed method, the plausible locations of a anchor point in the reference image are constrained by the epipolar geometry (see the depth sampling procedure described in section 9.5). This epipolar constraining strategy substantially reduces the number of plausible locations but requires a calibrated scenario to operate. This scheme is similar to the one proposed in DAISY Tola et al. [2010], but does not apply to the other methods. In order to perform fair comparisons, we introduce the epipolar constraining in the SoA solutions. This is achieved by restricting the plausible corresponding points in the reference image to the same set of points obtained by our depth sampling strategy. Therefore, all the methods are evaluated on the same depth hypotheses, thus on the same points on the reference image: $(\boldsymbol{\psi}'(\lambda_{min}), ..., \boldsymbol{\psi}'(\lambda_{\Pi}^{(h)}), ...\boldsymbol{\psi}'(\lambda_{max}))$. This scheme produces epipolar constrained versions of SIFT (SIFT-EC), LIOP (LIOP-EC) and SSID (SSID-EC) and is also used to define the plausible correspondences of DAISY. ASIFT and NRDC available codes do not allow to introduce this epipolar constraining, and the comparison with these methods would be restricted to the results included in Figure 9.10.

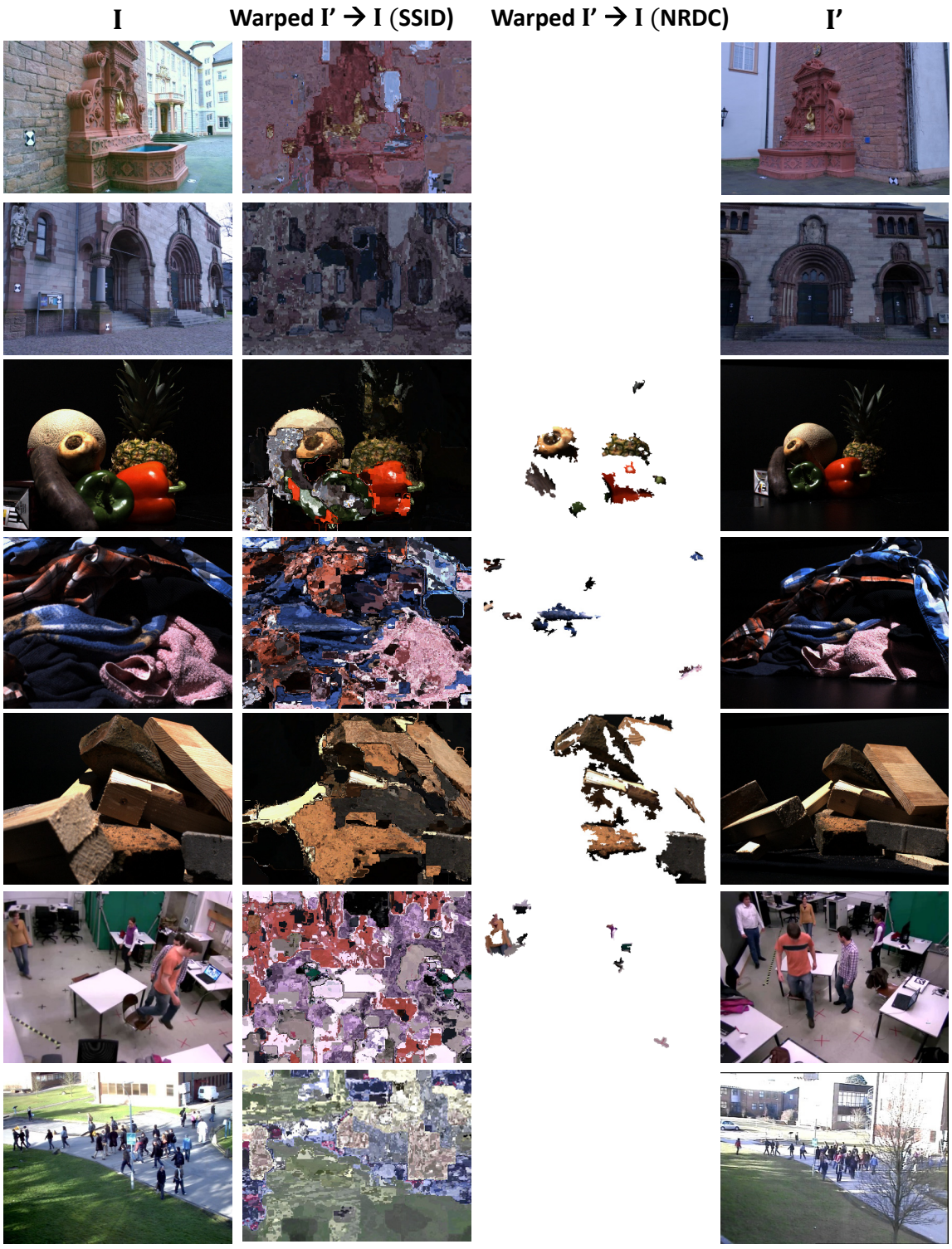


Fig. 9.10. Data-set complexity according to image warping by the correspondences obtained by SSID (through SIFT-flow) and NRDC. Whereas SSID always aligns the whole scene, NRDC just warp high confidence areas (the rest are leaved empty). See text for details and discussion.

SoA methods configuration

Default parameters have been used for all the evaluated methods. However, in order to adapt them to their epipolar constrained versions, slight modifications have been performed on their operation.

The potential of the SIFT and the LIOP descriptors relies partially on the scale-selection process performed at the point detection stage. As neither the anchor nor the corresponding points would be established by detection methods, a preliminary scale-selection process has been performed for these two methods. In particular, we construct the scale-space of the source and reference images with 200 scales by using a Gaussian kernel of 1.6 standard deviation. Then, we obtain the optimal scale for each point by maximising the absolute value of the difference of Gaussian operator (DoG) obtained for each scale.

A default $\gamma = 37.5$ is used for the SSID-EC method, as suggested by the authors in Trulls et al. [2013]. These so-configured descriptions are compared via the $l^2 - norm$ for each projection hypothesis as it is the metric used by all the authors.

Besides, we extract the DAISY description at the default configuration, aligning it with the epipolar lines and use the predefined masking scheme defined in Tola et al. [2010] to handle occlusions in the comparison stage.

Measuring performance

In order to perform fair comparisons, we shape a framework to evaluate local image descriptors on even grounds. To this aim we have manually selected points in the source image of each scene. Ground-truth correspondences on the other scene image were obtained through the depth maps (*fountain* and *herzjesu*), via the point-cloud (*greens*, *fabric* and *wood*) or by epipolar constrained manually annotation (*indoors* and *outdoors*).

The number of points selected for each scene varies between 30 and 70. The points are selected on potentially problematic image regions, including: feature-homogeneous areas; points which support is prone to be occluded; projected surfaces which present texture patterns which might be repeated somewhere else on the image; and points placed in projected surfaces which appearance strongly change between the two views. The annotated point and their associated ground-truth correspondences are included in the first two rows of Figure 9.12.

Our evaluation is independent of the detection stage. This conveys substantial benefits: any image point can be anchor (not just the singular points) and the influence of detection noise is eliminated. However, it also present a significant problem. Ground-truth annotations of the points have been performed based on either noisy information—depth maps were obtained by laser scanning, point-clouds by structured light—or by human annotation.

The problem, hence, is to define how close should be a corresponding point from its annotated ground-truth position to be considered part of a correct correspondence. To tackle this

uncertainty, we propose to establish a flexible criterion to measure the methods performance. In particular, we considered a correspondence correct if the matched point is placed at a spatial distance to its annotated position smaller than 10 pixels with independence of the image resolution. Furthermore, we use a rank-basis detection rate for quantitative evaluation (as in Simo-Serra et al. [2015]) . In particular, for the m-top ranked matchings the detection rate can be defined as:

$$\text{Detection Rate } (m) = \frac{100 \cdot \Upsilon_C(m)}{\Upsilon} \tag{9.24}$$

, where $\Upsilon_C(m)$ is the number of anchor points on the source image that have a correct correspondence among the top ranked m candidates in the reference image and Υ is the total number of anchor points on the source image.

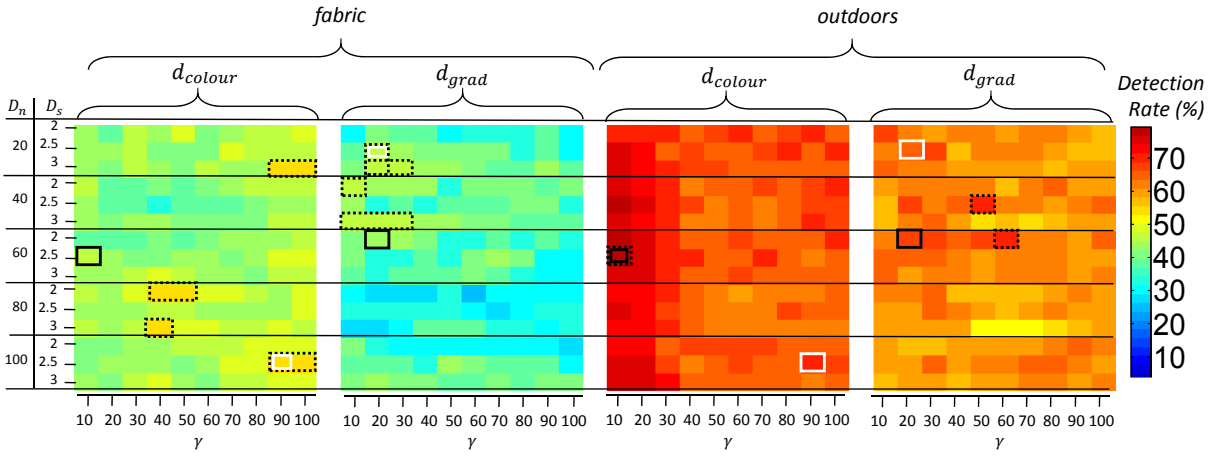


Fig. 9.11. Sensitivity analysis: Proposed method performance for different values of the parameters γ, D_S and $D_{\boldsymbol{n}}$. Average matching rates for *fabric* (left) and *outdoors* (right) scenes obtained using either colour-based (d_{colour}) or gradient-based (d_{grad}) distances for the 150 combinations of the parameters: $\gamma \in [10, 20, ..., 100]$, $D_S \in [20, 40, ..., 100]$ and $D_{\boldsymbol{n}} = [2, 2.5, 3]$. The matching rate is colour-coded (the redder the higher). Matching rates are obtained for $m = 10$. The best parameter configuration for each pair scene-distance is indicated by a black-dashed rectangle. The best configuration for each distance aggregating results for both sequences is indicated by a black-solid rectangle. A trade-off configuration which balances results according to the number of anchor points in each scene is indicated by a white-solid rectangle. See Table 9.4 for detection rates achieved for these three configuration. See Table 9.5 for the final selected parameters (from the trade-off scenario) and text for details and discussion.

Sensitivity analysis

In this section we study the influence of the method parameters in its performance and set their values. The method has four parameters, namely: the exponential decay of the weights

γ, the polar density sampling D_S, that defines the number of description points, the normal sampling $D_{\boldsymbol{n}}$, to sample the normal space, and the epipolar sampling Δ that controls the depth hypotheses sampling and the plausible locations on the reference image. We fix $\Delta = 1$, to obtain pixel-wise sampling of the epipolar line, lower values doubled the number of hypotheses and did not provide significant improvements.

In order to find the optimal values for the other three parameters we perform a sensitivity study of the parameters by analysing two scenes in the data-set: *outdoors* and *fabric*. These scenes have been selected as they constitute the extrema of the data-set in terms of the characteristics of the captured surfaces. On one hand, in *outdoors*, surfaces are captured at a small scale and most of the surface detail (texture) is lost. On the other hand, in *fabric*, surfaces are captured at a large scale, with prominent detailed patterns on different spatial frequencies which are also continuously repeated in other areas of the scene. Furthermore, the resolution of the images in *fabric* is more than two times that of *outdoors*, hence, the number of hypotheses for the *fabric* scene is significantly bigger than that in outdoors ($\sim 2D_{\boldsymbol{n}}$ times bigger). This increases the number of explored hypothesis and the likelihood of finding incorrect matchings.

We compute the Detection Rate ($m = 10$) for a wide range of values for each of these parameters leading up to 150 combinations of parameter values per comparison distance (d_{colour} and d_{grad}) and scene. Results of these processes are included in Figure 9.11. Let us discuss these results on three basics: operation of each scene, effect of the parameters and optimal parameter configuration.

Operation on each scene. On one hand, the operation of the proposed approach in the *outdoors* scene is substantially better than in the *fabric* scene (see performance for best configurations in Table 9.4). The presence of repetitive textures and the consideration of a higher number of hypothesis harms the operation in the *fabric* scene. On the other hand, the performance achieved by means of the colour-based distance d_{colour} slightly outperforms that achieved by the gradient-based distance d_{grad} in both scenes. However, for both distances, the proposed approach on its best-configured parametrisation is able to achieve detection rates over or around 50% (see 9.4), outperforming most of the state-of-the-art solutions (see Table 9.6).

Effect of the parameters. Results in Figure 9.11 suggest that the parameter values severely affect the performance of the proposed approach for the analysis of the *fabric* scene, whereas its effect on the analysis of the *outdoors* scene is less evident. According to equation 9.9, the higher is γ the lower is the weight associated to description samples in the anchor support. Hence, the higher is γ the lower is the contribution of neighbouring samples both in the appearance transformation scheme (see equations 9.17 and 9.18) and in the comparison stage (see equations 9.21 and 9.22). The operation on the *fabric* scene benefits from the use of high γ values probably due to the large scale of the objects and to the detail of their repetitive textures. Both factors turn the representativeness of slightly dissimilar samples in the anchor

Detection Rate (in %)	*fabric*		*outdoors*	
	d_{colour}	d_{grad}	d_{colour}	d_{grad}
optimal performance	53.13	46.88	79.41	69.12
optimal overall performance	46.88	43.75	79.41	67.65
trade-off performance	53.13	46.88	69.12	66.18

Table 9.4: Detection Rate for the sensitivity analysis (see Figure 9.11).

support less useful as they do not convey distinctive cues for description.

The influence of the polar density sampling parameter D_S on the approach performance does not present an structured operation pattern. The higher is D_S the lower the number of description samples in the anchor support. Hence, results suggest that using more samples for the description may be beneficial or detrimental in terms of the other two parameters. In any case, the best operations seems to be achieved for intermediate D_S values ($D_S = 2.5$).

Finally, the normal sampling D_n parameter is, by far, the most problematic. Different values of D_n not only convey different numbers of plane orientations, but also a different sampling of the orientations and, hence a different set of normal vectors. Large objects or closely captured objects in one of the views (as in the *fabric* scene) may be subjected to larger projective distortions than small objects. Small deviations of these sampled orientations from the *real* plane orientations of the scene can derive, for these larger captured objects, in enlarged erroneous estimations of the plane-induced homography. In our opinion, this is the main cause for the decrease of operation performance on the *fabric* scene. We plan to study this effect and alternative solutions to set D_n in our future work.

Optimal parameter configuration. The optimal performance of the algorithm for each scene and each distance is achieved for different configuration parameters (see black-dashed rectangles in Figure 9.11). Aggregating per point results (optimal overall performance) for all the scene on a distance basis derives in poor operation performances on the *fabric* scene (see black-solid rectangles in Figure 9.11 and detection rates in Table 9.4). The number of anchor points in the *fabric* scene (32) is substantially lower than the number of anchor points in the *outdoors* scene (68). Aggregating results but balancing on the different number of anchor points results in best achievable performance on the *fabric* scene at the expense of a slightly performance decrease in the *outdoors* scene (see white-solid rectangles in Figure 9.11 and detection rates in Table 9.4). We opt for this last configuration parameters as a trade-off solution. The rest of the results in this chapter are extracted by configuring the proposed approach with these parameters, which are included in Table 9.5.

Parameter	Value		Description
	d_{colour}	d_{grad}	
γ	90	20	weights exponential decay (sets influence of description samples in the anchor support)
Δ	1		epipolar line sampling (sets number of depth hypotheses)
D_S	2.5		polar sampling (sets number of description samples)
$D_{\boldsymbol{n}}$	100	60	normal hypotheses sampling (sets number and nature of the plane orientations explored)

Table 9.5: Configuration parameters of proposed method and values used in experiments. The values for γ, D_S and $D_{\boldsymbol{n}}$ are set by a sensitivity analysis (see Figure 9.11).

Results discussion

Table 9.6 includes the methods performance obtained for $m = 10$, i.e. accepting a matching as correct it the ground-truth position is among the ten first best-scored (associated with a lower matching distance) candidates. Results indicate that the proposed approach under its colour characterisation operates the second (after DAISY) when compared with epipolar constrained versions of top-performing local descriptions in the state-of-the-art. Furthermore, it is the only able to operate consistently (with detection rates over 50%) in every analysed scenario.The gradient version of the proposed approach operates slightly better than S-SID.

In order to also provide the reader with information about the ranking order of the correct matchings, we include in Figures 9.12-to-9.18 qualitative comparisons of the methods performance on the analysed scenes. In the Figures, it can be observed that the proposed approach, in its both configurations (colour and gradient) is together with DAISY the only one able to operate decently in all the evaluated scenarios. Additionally, it can be shown that the detected matchings usually appear among the top-ranked ones (associated to blue colour of the points). Finally, it is interesting to see that in several situations, the colour and the gradient failures are complementary, i.e. when one fail the other success. This is a clear motivation for a future combination of both measures.

The reasons why our algorithm operation is slightly behind that of DAISY are varied: the excess of hypotheses, the use of relatively simple features, the calibration inaccuracies, and the spatial margin. First, the proposed approach analysed a higher number of hypotheses than any of the other evaluated. In particular, we evaluated the other algorithms according to the depth part of the hypotheses, whereas the proposed method generates 100 (colour) and 60 (gradient) additional plane orientation hypotheses for each depth one. Anyway, the same ranking threshold ($m = 10$) has been used for all the schemes. Second, the simplicity of the characterisation features may harm the operation of our algorithm in some cases, specially when the appearance transformation is not lineal. Third, our comparison is quite sensitive to small displacements to the sampled surface orientations, which effect is specially relevant in the poorly calibrated scenes (*indoors* and *outdoors*). Fourth, DAISY, mainly relying on image gradients; hence operating

Scene (# points) / Method	SIFT-EC	DAISY-EC	LIOP-EC	S-SID-EC	Proposal (d_{color})	Proposal(d_{grad})
fountain (50)	16.00	72.00	44.00	78.00	70.00	64.00
herzjesu (50)	38.00	86.00	48.00	56.00	70.00	48.00
greens (39)	30.77	56.41	30.77	58.97	76.92	58.97
fabric (32)	12.50	37.50	12.50	34.38	53.13	46.88
wood (51)	11.77	76.47	21.57	78.43	72.55	68.63
indoors (72)	16.67	83.33	40.28	58.33	65.28	58.33
outdoors (68)	30.88	88.24	42.65	45.59	69.12	66.18
Total (362)	22.65	75.14	36.19	59.12	68.51	59.67

Table 9.6: Quantitative comparison of proposed method with the state-of-the-art. (EC stands for epipolar constrained) in terms of detection rate (in %) obtained for $m = 10$.

by differentiating between textured and not-textured areas, may be benefiting from the spatial range that was used to allow the match (10 pixels). This situation is specially relevant in the results obtained for the *outdoors* scene, where anchor points are mainly set over the people in the scene, which appear all together in small spatial area on the other image. As the rest of the image is little-textured (associated to small gradients) DAISY matchings in the people area are mostly considered correct, with independence of their real accuracy in the location of the anchor points.

9.7 Chapter conclusions

In this chapter we have developed a method to create and adapt anchor supports by means of the use of a fuzzy region segmentation and of a calibration constrained set of projective hypotheses. Through the proposed method we aimed to wide the scope application of wide-baseline point correspondence algorithm. In particular, the proposed method present a framework that hypothetically allows to match image points with independence of the scene surface orientation on which their three-dimensional projection lies under two assumptions: (i) the surface can be locally approximated by planes and (ii) images of the surface are, at least, partially projected on the two views. Obtained results in challenging scenes are promising. Our future work will be devoted to explore the use of alternative characterisation schemes and to extrapolate this scheme to dense-matching schemes and multi-camera scenarios (scenarios with more than two-views).

Fig. 9.12. Qualitative comparison of proposed method with the state-of-the-art. (fountain). First row: spatial position of anchor points (left), ground-truth positions in the other view (right). Second row: Epipolar constrained SIFT (left) and DAISY (right). Third row: Epipolar constrained LIOP (left) and S-SID (right). Fourth row: Proposed method based on colour comparison (left) and based on gradient magnitude comparison (right). Ground-truth points are represented by white circles. Rank of the correspondence is indicated by a colour code from 1^{st}(dark blue) to 10^{th} (red). Lower ranked correspondences are replaced by the ground-truth position to indicate a failure of the algorithm. See text for discussion.

Fig. 9.13. Qualitative comparison of proposed method with the state-of-the-art. (fountain). First row: spatial position of anchor points (left), ground-truth positions in the other view (right). Second row: Epipolar constrained SIFT (left) and DAISY (right). Third row: Epipolar constrained LIOP (left) and S-SID (right). Fourth row: Proposed method based on colour comparison (left) and based on gradient magnitude comparison (right). Ground-truth points are represented by white circles. Rank of the correspondence is indicated by a colour code from 1^{st}(dark blue) to 10^{th} (red). Lower ranked correspondences are replaced by the ground-truth position to indicate a failure of the algorithm. See text for discussion.

Fig. 9.14. Qualitative comparison of proposed method with the state-of-the-art. (fountain). First row: spatial position of anchor points (left), ground-truth positions in the other view (right). Second row: Epipolar constrained SIFT (left) and DAISY (right). Third row: Epipolar constrained LIOP (left) and S-SID (right). Fourth row: Proposed method based on colour comparison (left) and based on gradient magnitude comparison (right). Ground-truth points are represented by white circles. Rank of the correspondence is indicated by a colour code from 1^{st}(dark blue) to 10^{th} (red). Lower ranked correspondences are replaced by the ground-truth position to indicate a failure of the algorithm. See text for discussion.

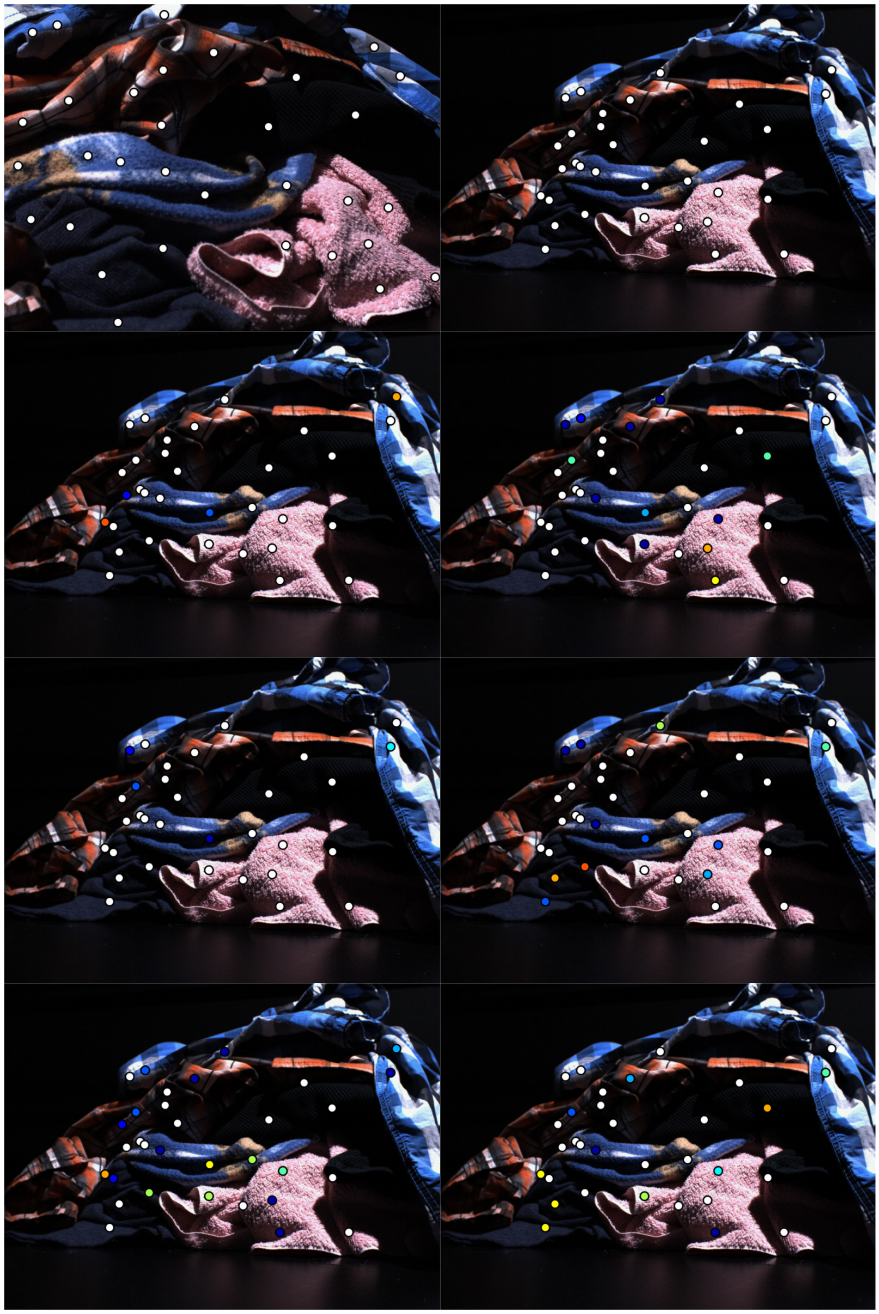


Fig. 9.15. Qualitative comparison of proposed method with the state-of-the-art. (fountain). First row: spatial position of anchor points (left), ground-truth positions in the other view (right). Second row: Epipolar constrained SIFT (left) and DAISY (right). Third row: Epipolar constrained LIOP (left) and S-SID (right). Fourth row: Proposed method based on colour comparison (left) and based on gradient magnitude comparison (right). Ground-truth points are represented by white circles. Rank of the correspondence is indicated by a colour code from 1^{st}(dark blue) to 10^{th} (red). Lower ranked correspondences are replaced by the ground-truth position to indicate a failure of the algorithm. See text for discussion.

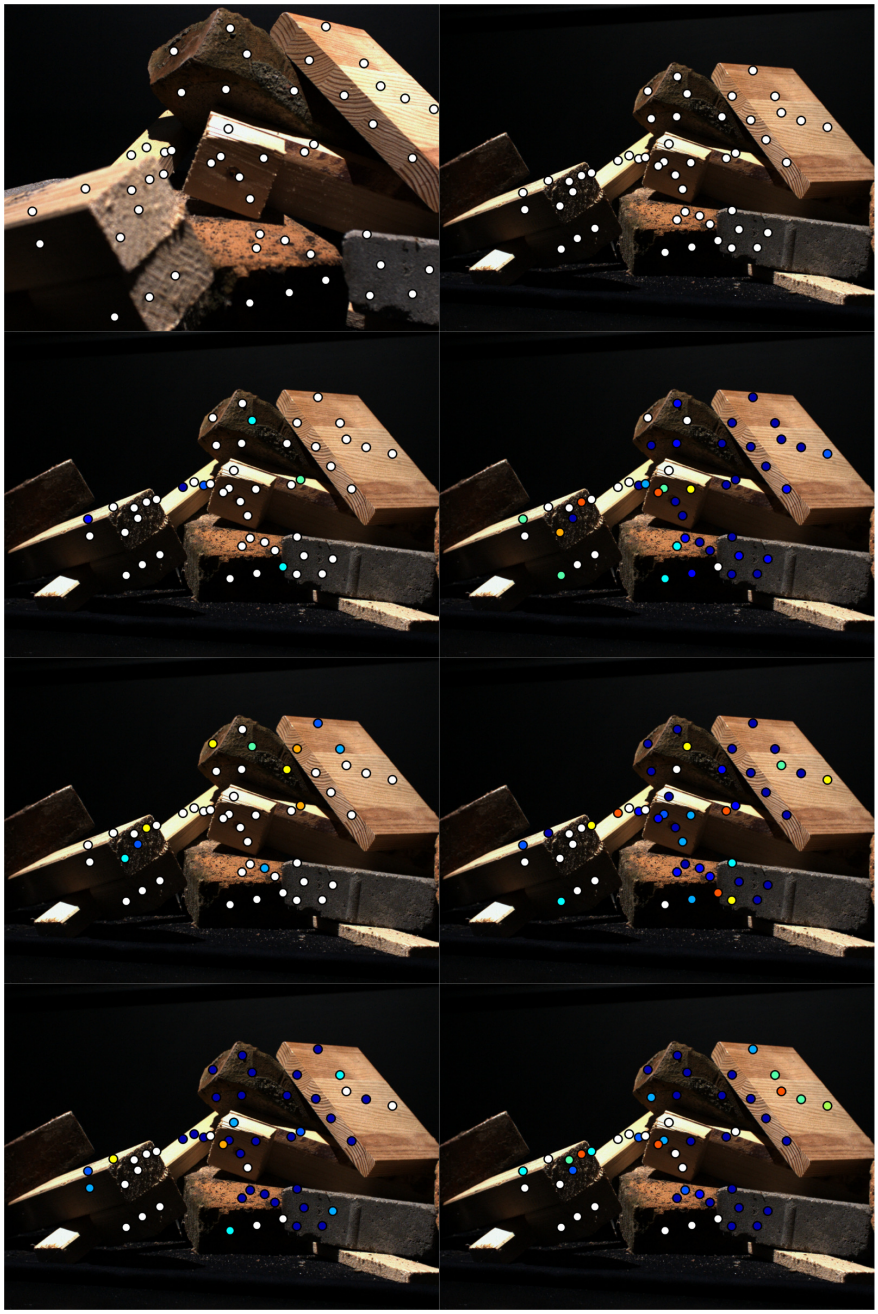


Fig. 9.16. Qualitative comparison of proposed method with the state-of-the-art. (fountain). First row: spatial position of anchor points (left), ground-truth positions in the other view (right). Second row: Epipolar constrained SIFT (left) and DAISY (right). Third row: Epipolar constrained LIOP (left) and S-SID (right). Fourth row: Proposed method based on colour comparison (left) and based on gradient magnitude comparison (right). Ground-truth points are represented by white circles. Rank of the correspondence is indicated by a colour code from 1^{st}(dark blue) to 10^{th} (red). Lower ranked correspondences are replaced by the ground-truth position to indicate a failure of the algorithm. See text for discussion.

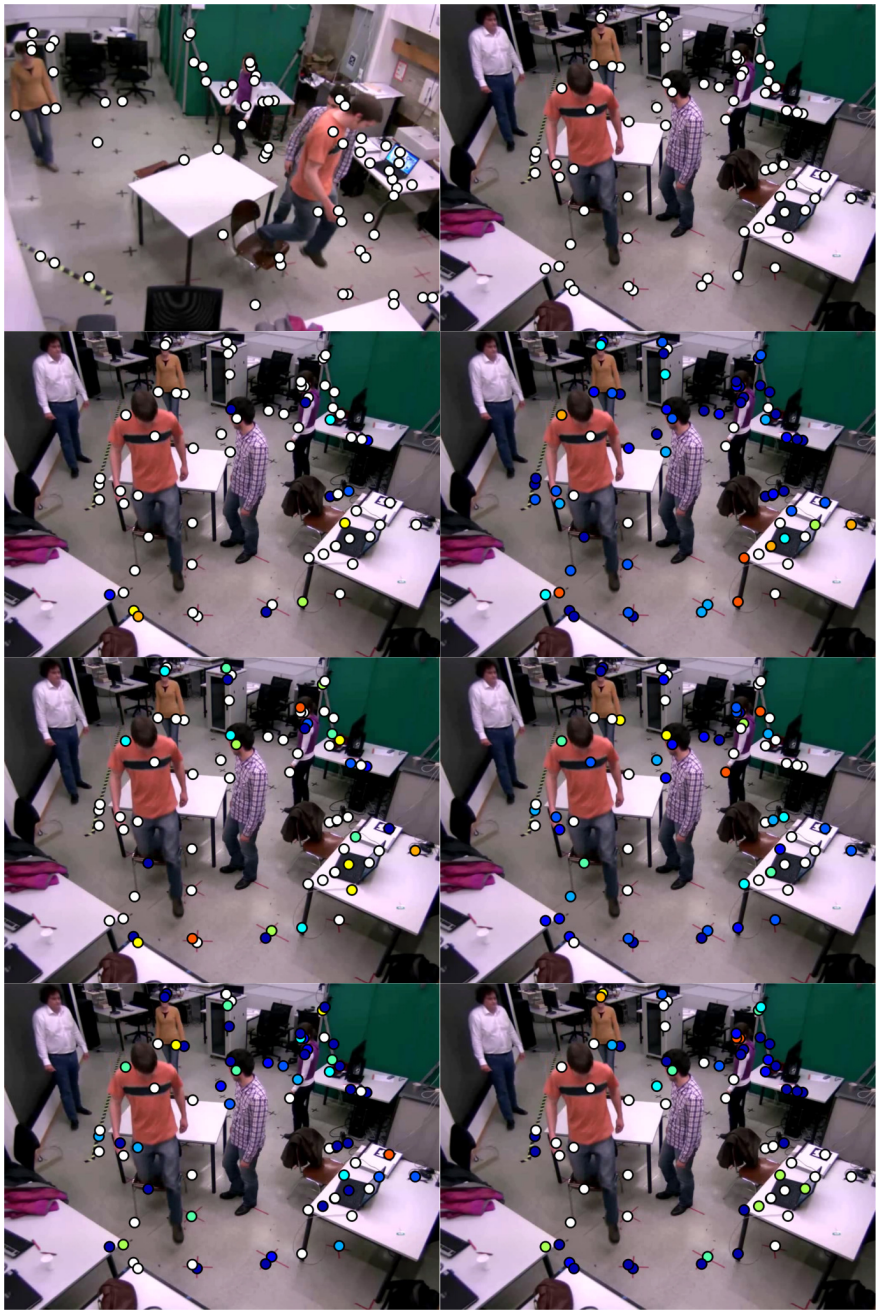


Fig. 9.17. Qualitative comparison of proposed method with the state-of-the-art. (fountain). First row: spatial position of anchor points (left), ground-truth positions in the other view (right). Second row: Epipolar constrained SIFT (left) and DAISY (right). Third row: Epipolar constrained LIOP (left) and S-SID (right). Fourth row: Proposed method based on colour comparison (left) and based on gradient magnitude comparison (right). Ground-truth points are represented by white circles. Rank of the correspondence is indicated by a colour code from 1^{st}(dark blue) to 10^{th} (red). Lower ranked correspondences are replaced by the ground-truth position to indicate a failure of the algorithm. See text for discussion.

Fig. 9.18. Qualitative comparison of proposed method with the state-of-the-art. (fountain). First row: spatial position of anchor points (left), ground-truth positions in the other view (right). Second row: Epipolar constrained SIFT (left) and DAISY (right). Third row: Epipolar constrained LIOP (left) and S-SID (right). Fourth row: Proposed method based on colour comparison (left) and based on gradient magnitude comparison (right). Ground-truth points are represented by white circles. Rank of the correspondence is indicated by a colour code from 1^{st}(dark blue) to 10^{th} (red). Lower ranked correspondences are replaced by the ground-truth position to indicate a failure of the algorithm. See text for discussion.

Part V

Part V. Conclusions and future work

Contents

This part concludes the Thesis by discussing on the personal motivation of the solutions proposed along the document and by reviewing the key contributions described up to this point.

"O fim de uma viagem é apenas o começo de outra."

José Saramago. (Viagem a Portugal, 1981)

Chapter 10

Achievements, conclusions and future work

10.1 Overall discussion on the strategies in the document

Along the Thesis presented in this document, as in the development of any research work, several decisions have been taken. We made these decisions in consonance with the faced problem and according to the state-of-the-art. However, voluntarily or involuntarily, these decisions are biased by our understanding of computer vision. Let us review some of the factors that partially defined our studies.

Machine learning vs semantic processing

In one of his less referenced studies (Shannon [1988]), C.E.Shannon explores two alternatives for programming a computer to play chess. The first of them, that he named the Type A strategy, consists in—given an opponent movement—exploring all the possible movements and all the possible consequences of such movements: creating a selection net. Then, the best movement is selected in consonance with this exploration. The second strategy, Type B, involves exploring only a subset of important or relevant movements as those that constitute *interesting* branches in the selection net, where the *interesting* concept is determined by the programmer experience or by a set of semantic rules. Nowadays we have generalised these solutions to many other problems ahead for chess. Methods which follow strategies framed as Type A are sometimes criticised to be too *machine learning processes.* The opposite strategies, e.g. those that rely on strong selective constraints on the *decision space* are claimed to be excessively *semantically designed processes.* Shannon argued for a solution of Type B to increase the computational efficiency of the system and to allow natural interaction with the human player. On one hand, recent and continuous developments of computational resources diminish the relevance of this premise. On

the other hand, the complexity of the problems faced by recent solutions are sometimes less constrained than chess movements; hence these problems are assumed to construct a larger and more branched decision net.

The solutions proposed in this Thesis may be classified as *semantically* designed. Besides chapter 8, none other solution proposed along this document relied on a machine learning scheme—albeit training data has been used to evaluate the sensitivity to key parameters of some methods—. We are strong defenders of *semantic* selection, but we are quite aware of the complexity in predetermining all the possible operation processes that the analysis of video data can entail. Our future work will explore the use of well-developed solutions to automatically discover relations amongst video data. Specifically, we aim to explore the potential capabilities of Convolutional Neural Networks.

Invariance vs Adaptation

This discussion was explicit in chapter 9; however, splitting hairs, we can extend it to the rest of the faced problems.

In video and image *understanding* approaches, the final target is to provide alternative representations of the content to ease the detection of objects, of their interactions or of the events happening on the scene. To this aim, researchers usually rely on pre-defined or pre-trained models to represented the anchor cue. The invariance vs adaptation discussion arises in the design and use of these models. Invariant design refers to the models which present a high robustness to the potential changes that may affect a modelled cue. Adaptable design entails that the model is flexible and can be adjusted to these potential changes. Throughout this document we have proposed approaches inspired by both designs. The region-segmentation methods in Part II relied on invariant descriptions to group pixels. Our aim was there to design invariant features able to handle inter-pixel variations. The first background subtraction solution in chapter 7 was designed on an invariant basis. There, we presented a BS refinement method which provides moderate robustness to illumination changes via the definition of illumination-blind (invariant) features. On the contrary, the additional contributions of the second solution were of an adaptation nature. In particular, the region-driven background model there proposed was able to adapt to the background dynamism. The local description methods presented in Part IV can be also described in these terms. On one hand, in chapter 8, we combined the use of invariant descriptions with a knowledge-modelling scheme to adapt the object identification to the number of visible parts of an object instance. On the other hand, in chapter 9 we presented an adaptable solution to geometrically and spectrally adapt an anchor support.

In our opinion, if the potential changes that may affect a cue are known or can be defined, invariant design is the best option. On the contrary, if the potential changes are unconstrained or unpredictable, adaptable design is required.

Holistic vs Local

We moderately explored the use of holistic information. For instance, in chapter 3 we divided existing region-segmentation approaches into global and local. Additionally, in chapter 4 we started from an analysis of the image data distribution to derive global constraints on the pixel grouping. However, aside from these chapters, global information has not been explicitly explored in any other chapter. We aim to change this in our future work. In our opinion, recent studies in the areas of context modelling (Torralba [2003]; Gijsenij and Gevers [2011]; Anand et al. [2012]; Choi et al. [2012]) suggest that if the overall scene configuration is recovered, it may constitute a key constraining cue for analysis. This was also partially suggested in this Thesis by the hypotheses constraining scheme of chapter 9.

Is it worth using regions?

This is probably the most relevant question in the Thesis, as all the document builds on it. Along the document we have provided some theoretical and experimental examples that motivate the use of regions. According to them, we believe that a preliminary region-segmentation of the image is beneficial for the majority of the methods—if the computation cost is not an issue—. Nevertheless, the potential benefits of the regions are conditioned to their proper extraction.

As any other analysis task, region segmentation is subject to processing inaccuracies. Errors made in low-level stages of analysis are usually corrected or refined in later stages. On the contrary, there is usually no turning back for region-segmentation errors. If, during region segmentation, some of the relevant scene contours are not detected, the semantic entities defined by these contours can be hardly recoverable by posterior stages of processing. This is the main motivation for the conservative design of the proposed solutions. For instance, in chapter 4, recall in boundary detection was prioritised over precision. Similarly, the use of a pixel-based background subtraction was required as a complementary method to the region-based solutions in chapter 7. With the same aim, in chapter 8, region segmentation was carried out at different coarseness levels to enhance the probability of achieving representative object parts. Finally, in chapter 9, a fuzzy region segmentation scheme was used in order to avoid the absolute assignment of pixels to regions, hence allowing the proposed scheme to handle with uncertainties derived from the data and the segmentation scheme.

10.2 Summary of achievements and main conclusions.

This document has been structured in five parts (being this the fifth part). Let us review the conclusions of each of the first four parts. The conclusions associated to the objectives stated in section 1.2 of chapter 1 are highlighted in bold.

Part I

Part I was devoted to introduce the studies in the document and specially to motivate the use of regions for image and video analysis.

To this aim, in chapter 1, we first started by motivating *the region* from three different point-of-views. We first placed the region as an intermediate processing analysis unit. The analysis results of computer vision approaches are usually grouped for post-processing; hence, changing the order of operation seemed a natural step: first grouping the pixels and then analysing these groups. From this ordering arose two different description spaces: the decision space (to group pixels into regions) and the feature space (on which to perform the analysis). We also claimed that, sometimes, the grouping process can derive into better statistical results. Finally, we connected region-based processing with human perception theories that, apparently, rely on the grouping of individual stimuli to drive object perception processes. The chapter continues by identifying the main objectives of the Thesis—which accomplishment will be evaluated here—and by enumerating its major contributions. The chapter ends with the presentation of the document organisation, including alternative reading orders and a description of the temporal evolution of the developed research. We believe that such description may help to understand the diversity of the research paths that have been explored along this document.

In chapter 2, the use of regions is further motivated. The chapter started with **a definition of the region concept**. To this aim, we reviewed definitions of the word *region* available in different online English dictionaries. These definitions were merged together to coalesce into a generic, but closed, definition of the region and of the region segmentation process. Then, a set of basic characteristics of the regions were presented: the region label, the region representative, the boundary of a region and the adjacency relationship between two regions. These characteristics were considered part of the region segmentation process because they inherently result from the region partition. The chapter continued with a review of the challenges that have classically motivated the use of regions in this document. In particular, the use of regions is motivated by their ability to: narrow the semantic gap, de-noise signals, sample the feature space, provide automatic adaptable support and—if extracted fuzzily—codify inter-pixel similarities.

Aside for their motivational aim, Part I was devoted to present the common concepts that are used along the document. Furthermore, Part I also sketched our understanding of the region and provided a basic structure on which the rest of the document is sustained.

Part II

Part II was focused in the region segmentation stage (RS).

Part II started with chapter 3, on which we reviewed and **organised top-relevant region-segmentation approaches**. The chapter began with a discussion on the problematic associated

with the evaluation of region segmentation techniques. There it was shown that humans perceived different regions when they were asked to segment a given image, hence, establishing a common criterion on how a good region segmentation is claimed to be an almost infeasible task. Then, the chapter continued with a description of the relevant factors that characterise a region segmentation approach: the features used for description, the management of edges and contours, the order of processing and the scale on which regions are extracted. Next, the proposed organisation was presented. We organised existing approaches by a dual-scheme. First, on an stage basis, in which five stages (some of them mandatory, some others optional) were defined: pre-processing, feature extraction, local analysis, globalisation and regionalization. Then, regarding the level of processing, distinguishing between local, global and combined approaches. This last organisation was used to arrange the selected approaches, whereas the stage-based organisation was implicitly used to describe each of these approaches. Five broad categories were finally set. Three of them were considered pure-local approaches: clustering, region-merging and mode-seeking. Just one of them was understood to be enough to explain pure-global approaches: energy-minimisation. Combined approaches were tagged as graph-based, a category which was further subdivided into hierarchical and contour detection approaches. The chapter ends by describing three of the existing data-sets for evaluation and the evaluation metrics used to determine (to some extent) the goodness of a region segmentation approach.

Chapter 4 proposed a novel integration of the Mean-Shift (MS) technique and the scale-space theory. In particular it started by reviewing both topics in detail, in order to motivate the proposed approach and to theoretically reinforce the design decisions made along the chapter. Then, **a strategy to automatically set a variable spectral bandwidth for every sample in a MS process** was presented. The strategy, designed to operate on discrete data distributions, was inspired by the scale-space theory and benefits MS by inhibiting its stagnation in plateau areas of the distribution; by accelerating its local mode-seeking processes; and by avoiding its convergence to non-global modes. Additionally, the strategy avoids the requisite of selecting the spectral bandwidth parameter, which usually determines the operation of MS approaches. However, the operation of the proposed approach, MS-RS, is still function of a threshold parameter. This threshold determines the minimum mode size (the number of samples assigned to a mode) for the mode to be considered a global mode. The use of the proposed approach for RS led to disparate results depending on the data-set analysed. On one hand, on a moderately textured data-set, the algorithm was able to achieve tightened-to-objects segmentations while respecting the fine details of the image. In contrast, the operation of the most-used MS approach in this scenario was shown to be strongly dependent of the bandwidth parameter. On the other hand, for a highly-textured data-set, the MS-RS still respected the object boundaries but severely over-segmented the image. A post-processing approach based on a hierarchical region-merging procedure was proposed to handle this problem. The merging-procedure relied on distances

between colour descriptions of the regions in the CIELab colour space to merge adjacent regions under several coarseness hypotheses. These hypotheses were established by an analysis of the colour distances distribution so that a generic set of hypotheses can be defined for images of different nature. Results indicated that the region-merging procedure, whereas benefited the operation of MS-RS to some extent, was still operating quite below the leading approach in the state-of-the-art. We concluded that the use of colour information was not enough to handle textured areas.

Chapter 5 was devoted to define a scheme to describe local-variability around a pixel. The chapter began with a review of existing methods for describing local variability emphasising Texton-based discriminative methods. We discussed the theoretical incongruities of these methods in order to present **the Discrete Cosine Transform (DCT) filter-bank as a suitable and easily configurable (fully defined by a single parameter, the block size) set of filters to describe local-variability**. By means of the block size, the nature, the scale and the number of total filters in the filter-bank is automatically defined. Additionally, the DCT has the ability to condense the majority of the information in the responses (coefficients) of a small subset of these filters. We took advantage of this ability and proposed a method to automatically select this subset of relevant filters for every pixel based on the representativeness of their responses. By means of a sensitivity analysis on 200 images, we arose to the interesting result that the number of relevant filters can be set the same for different images if the responses of the filters are arranged according to their representativeness and small errors are tolerated. This study allowed us to discard a substantial number of non-relevant filters; however, it entailed a problem in the comparison of the responses of these filters. As the filters might be differently selected for different image pixels, the straight comparison of their responses required a preliminary comparison of the filter themselves. To cope with this problem, we defined a measure to compare any two filters in the DCT filter-bank. The measure, albeit inspired by subjective premises, fulfilled all the conditions to be called a metric. Building on this metric, we derived an enhanced metric that also included response intensity in the comparison and accounted for multi-scale processing. The so-built metric was used to create a contour map on which image pixels were assigned a likelihood of being part of a contour. Preliminary results suggest that the scheme may be useful to define local-variability transitions.

The huge amount of existing RS approaches discourage new proposals on the field. However, we believe that most of the problems commonly associated with regions are consequences of inaccuracies committed during their procurement. We face RS from a thorough perspective and study the problematic of classical, yet successful, schemes. In chapter 4 we remove the dependency of MS to its most problematic parameter. This is achieved at the expense of introducing a new parameter. Whereas we think that this parameter is easier to set properly, further research is required. MS is a pure bottom-up approach; our proposal turns MS into

a combined approach that relies on global information to drive local analysis. In chapter 5 we faced a completely different approximation; instead of grouping pixels we search for scene transitions. These transitions are searched by studying changes on the local description of pixels. Top-performing existing solutions are based on the application of a set of crafty designed filters but do not consider the relationships among these filters when comparing their responses. The foundational cosine functions through which the DCT is generated provide straight associations among the filters that compose the DCT filter-bank. Hence, we considered the DCT as a simple point on which starting our research on filter inter-similarities. Anyway, we plan to further explore these relations in alternative filter-banks in our future work and to test the proposed solution on different scenarios.

Part III

Part III addressed the use of regions as complementary analysis units for background subtraction (BS).

Chapter 6 reviewed **recent and top-relevant approaches in the field of BS**. After a definition of the problem, the chapter started with a description of the challenges that a BS approach should face. Relevant approaches were organised on a per-stage basis relating these processing stages with the faced challenges. This organisation helped to bring out the strengths and weaknesses of BS approaches ahead of their quantitative evaluation. The problems of this evaluation were discussed, highlighting two of them, derived from the existence of a single proper evaluation data-set: results over-refinement (which affects the shape of the foreground map to increase results precision) and *ad hoc* design of the methods (which inhibits the study of challenges poorly represented in the data-set). The chapter ended by describing the evaluation metrics in BS and with an explanation of the scarcity of region-based solutions in the state-of-the-art of BS. In particular, we concluded that the computational cost of region segmentation approaches hinders their use for BS.

In spite of the previous conclusion, and under the expectancy of future improvements in hardware and software processing, chapter 7 presented **two region-based solutions which results suggested that region analysis can benefit pixel-based BS**. The first solution relied on a very simple MS approach to derive illumination-blind regions so that which the effect of illumination artefacts (shadowed and over-lit areas) in the RS was substantially reduced. The designed MS (germinal of the approach described in section 4) operates by adapting the kernel to search for albedo continuity and by merging resulting reflectance-homogeneous regions by searching for the alignment of their RGB colour vectors. Overall, the scheme was proven to be effective in extrapolating pixel-level results from correctly classified areas to (connected) incorrectly classified ones. The second solution enhanced this one by also considering the temporal evolution of the region. To this aim, a novel region-driven background model was proposed.

The technique relies on a covariance-based modelling of a configurable set of features. An eigenvalue comparison was then proposed to detect not-modelled foreground samples. The model was defined of a multi-layer nature such that it can cope with temporal variations of the region. Preliminary results suggested that the method was able to achieve tightened-to-foreground masks without relying on any post-processing approach.

These region-driven approaches are complemented by Appendixes A and B. Appendix A described a feasibility study on the use of the DCT-based characterisation and associated metric defined in chapter 5 for increasing foreground-background separability. Appendix B proposed a multi-layer and multi-class background model which relied on class-driven temporal inhibition mechanisms to avoid the corruption of the background model by wrongly classified foreground samples.

Part III and associated Appendixes presented our contributions to BS. Whereas the experimental results of these solutions were promising, further evaluation is required to derive solid conclusions. The use of region-based contributions to BS is hindered by the real-time requirements of most BS applications. Furthermore, in our opinion, there is still a lack of research in multi-class BS. Future work will be devoted to exploit contextual information in order to generate relevant/irrelevant maps that can be used to define a subset of pixels on which region processing can be performed / skipped. We believe that, if the computational complexity of region segmentation is reduced by diminishing the number of pixels to group, the use of region-based approaches can provide substantial improvements on the quality of pixel-based BS.

Part IV

Part IV deals with the use of regions for local description (LD) of image points.

Chapter 8 begins with a motivational section that aims to establish links between theories of human perception and computer-vision approaches for object identification. In that section we motivate an alternative approach for object identification that was inspired by human perception. To this aim, we suggest that, differently to the majority of the solution proposed by state-of-the-art approaches, nor a huge set of training samples nor a set of objects-specific models are required for identification. On the contrary, we hypothesised that knowledge can be generalised from a small set of samples and that a single model can be used to store this knowledge. On one hand, knowledge generalisation was achieved by partitioning training object instances at several level of coarseness, **generating multi-coarse part-wise descriptions of an object; thereby achieving robustness to object occlusions by hypothesising on the expected visible parts of an object**. Furthermore, to cope with changes in the capture point-of-view, parts were aligned by the orientations of the singular points they contained. For the LD of these parts we proposed two novel description strategies that relied on the region-masking of successful state-of-the-art two-dimensional and three-dimensional descriptors. On the other

hand, knowledge storing was driven by the distributed encoding of the LD in a neural model and objects were identified by measuring the responses to this model. Results were extracted under the assumption that objects had been previously segregated from a severe occlusion scenario, albeit presenting very different appearances to those observed during their training. Under these premises, results suggested that the proposed scheme was able to outperform the operation of a top-performance state-of-the-art LD for the task of object identification. Our future work will be devoted to avoid the requirement of a preliminary object segregation stage.

Chapter 9 proposed a completely different alternative to match points across images. The solution **uses the scene calibration to constrain the possible geometric deformations across images of a region-based LD support of an anchor point**. These possible deformations are modelled by the using of plane induced homographies. The projections of the anchor support obtained through these homographies represent the candidate supports. The likelihood of these candidates being the real projections of the anchor support is obtained by comparing them, either by colour or gradient LD, with the anchor support. Fuzzy regions are used to define the extent of the support and to weight the influence of its samples in the description, according to their similarity to the anchor point. Furthermore, in order to also cope with appearance transformations of the support, these fuzzy regions are also used to weight a linear feature transformation of the support LD. Results, extracted on complex scenarios, are promising.

The contents in this part are inspired by the idea proposed in Tola et al. [2010] of using inhibition masks to confront occlusions. Using regions instead of predefined patterns allows to adapt the inhibition to the image content. When we derived our first solution on this topic (see Navarro et al. [2014] and Appendix C), we were not aware that a very similar approach was previously designed (Trulls et al. [2013]). In chapter 8 we apply this adaptable masking scheme to two widely used description methods. In chapter 9 we instead use the idea proposed in Trulls et al. [2013]. This description-inhibition field, is in fact a hot research topic nowadays, so that there are plenty of potential lines of research. A substantial part of our future work will be devoted to explore some of these lines.

Chapter 11

Hitos, conclusiones y trabajo futuro

11.1 Discusión global sobre las estrategias seguidas a lo largo del documento

Durante el desarrollo de la Tesis presentada en este documento, como suele ocurrir durante el desarrollo de cualquier trabajo de investigación, se han tomado una serie de decisiones en función del problema estudiado y en consonancia con el estado del arte relacionado con dicho problema. Sin embargo, ya sea voluntaria o involuntariamente, existe un sesgo derivado de nuestra comprensión o interpretación de la tarea de visión por computador. A continuación describiremos algunos de los factores que definen parcialmente nuestra forma de enfocar esta tarea.

Aprendizaje máquina frente a procesamiento inspirado en reglas semánticas.

En uno de sus estudios menos referenciados, Shannon [1988], C.E.Shannon describe dos alternativas de diseño para programar un ordenador *capaz* de jugar al ajedrez. La primera de ellas, denominada estrategia tipo A, consiste en,explorar todos los posibles movimientos a realizar y sus consecuencias a partir de un movimiento del oponente. Este proceso deriva en la creación de una red de selección. Con la red construida, el mejor movimiento se selecciona considerando la red al completo. La segunda estrategia, denominada tipo B, implica la exploración de sólo un subconjunto de movimientos, que podrán entenderse como ramas de interés dentro de la red de selección. El concepto de interés viene determinado por la experiencia del programador o por determinadas reglas semánticas prefijadas. En la actualidad, estas estrategias se han aplicado a diversos problemas más allá del ajedrez. Aquellos métodos que pueden ser considerados como de tipo A son habitualmente criticados por estar excesivamente orientados al aprendizaje máquina, es decir, que exploran hipótesis que son plausibles pero, en general, carecen de sentido (Por ejemplo intercambiar peón por dama sin obtener ventaja posicional alguna). Las estrategias

opuestas, es decir, aquellas que radican en el uso de restricciones de selección estrictas, son a veces catalogadas como demasiado semánticas o demasiado guiadas por nuestro conocimiento del problema. Siguiendo el ejemplo anterior, la ventaja posicional podría adquirirse tras una larga sucesión de movimientos derivados del intercambio, difícil de predecir por un jugador de nivel medio. Shannon optó por una solución de tipo B con el fin de incrementar la eficiencia computacional del algoritmo, permitiendo así una interacción natural con un hipotético usuario humano. Por un lado, los continuos avances en los recursos de computación disminuyen la relevancia de la premisa seguida por Shannon. Por otro lado, los problemas afrontados en la actualidad son, en algunos casos, más complejos que el ajedrez, por estar menos constreñidos. Por ende, la red de selección de estos problemas suele ser más profunda y ramificada que la creada para el ajedrez.

En general, las soluciones propuestas en esta Tesis pueden ser clasificadas como de inspiración semántica. Con la excepción del capítulo 8, ninguna otra solución requiere de un aprendizaje máquina clásico para operar, aunque sí se han utilizado conjuntos de entrenamiento para evaluar la sensibilidad de las soluciones propuestas a parámetros clave. En definitiva, si bien podríamos catalogarnos como firmes defensores de la selección semántica, comprendemos la complejidad de predeterminar todos los potenciales procesos de operación que implica el análisis de vídeo. Por ello, nuestro trabajo futuro explorará el uso de soluciones bien fundamentadas que permiten descubrir o listar automáticamente relaciones entre los datos de análisis. Específicamente, nuestros esfuerzos se centrarán inicialmente en la evaluación de las capacidades de las Redes Neuronales Convolucionales (CNN por sus siglas en inglés).

Invarianza frente a adaptación

Esta discusión, explícita en el capítulo 9, puede extenderse al resto de los problemas estudiados. En las aproximaciones de interpretación automática de imágenes y vídeo, el objetivo último puede entenderse como el de suministrar representaciones alternativas del contenido que faciliten la detección de objetos, de las interacciones entre objetos o de cualquier tipo de evento catalogable que tenga lugar en la escena capturada. Con este fin, suelen utilizarse modelos predefinidos o pre-entrenados para representar el objeto de búsqueda. La discusión entre invarianza y adaptación radica en el diseño y uso de estos modelos. Los diseños de tipo invariante establecen modelos robustos a potenciales variaciones que pueden afectar a la característica modelada. Por el contrario, los diseños adaptables implican que el diseño del modelo es flexible y puede ajustarse a estas variaciones. A lo largo del documento hemos propuesto enfoques inspirados en estos dos tipos de diseño. Los métodos de segmentación en regiones descritos en la Parte II utilizaban descriptores invariantes para agrupar píxeles. Para ello, utilizaban características robustas frente a pequeños cambios de apariencia entre píxeles. La primera solución en el área de sustracción de fondo (capítulo 7) también se diseñó bajo un enfoque invariante. En particular,

se propuso un método de refinamiento que proporcionaba robustez moderada a los cambios de iluminación mediante el uso de características pseudo-invariantes a dichos cambios. Por el contrario, las contribuciones de la segunda aproximación son de naturaleza adaptativa, enfocadas principalmente a adaptar las descripciones a los cambios producidos por el dinamismo del fondo. Los métodos de descripción local definidos en la Parte IV del documento también pueden ser descritos en términos de invarianza y adaptación. Por un lado, en el capítulo 8, proponemos combinar el uso de descriptores invariantes con un modelo de conocimiento que permite adaptar el proceso de identificación de objetos al número y a la naturaleza de las partes visibles de una instancia del objeto a identificar. Por otro lado, en el capítulo 9, presentamos un esquema para adaptar geométrica y espectralmente el entorno de descripción del punto buscado.

En nuestra opinión, si los cambios potenciales que pueden afectar a una determinada característica están predeterminados o pueden, al menos, definirse, el diseño invariante es la mejor opción. Por el contrario, si éstos son indeterminados o impredecibles, el uso de un diseño adaptativo es altamente recomendado.

Información global frente a información local

Hemos explorado moderadamente el uso de información global. Por ejemplo, en el capítulo 3 dividimos las estrategias para la segmentación de regiones existentes entre globales, combinadas y locales. Además, en el capítulo 4 utilizamos un análisis de la distribución global de los datos para establecer restricciones sobre el proceso de agrupación de píxeles en regiones. En cualquier caso, el estudio de la información global se ha limitado a estos capítulos. Nuestro objetivo es cambiar el enfoque en el trabajo futuro. En particular, los estudios relativamente recientes en el área de modelado del contexto (Torralba [2003]; Gijsenij and Gevers [2011]; Anand et al. [2012]; Choi et al. [2012]) sugieren que si la configuración global de la escena está disponible, ésta puede suponer una constricción excelente para el análisis, como se sugiere implícitamente en el proceso de restricción de hipótesis descrito en el capítulo 9.

¿Merece la pena utilizar regiones?

Esta es, probablemente, la pregunta más relevante de esta Tesis, dado que todo el documento se sustenta en ella. A lo largo del documento hemos suministrado ejemplos teóricos y experimentales para motivar el uso de regiones. Utilizándolos como argumento, creemos que una segmentación de la imagen en regiones, preliminar al análisis, es beneficiosa para la mayoría de aproximaciones, siempre que el coste computacional no sea una demanda prioritaria. En cualquier caso, los beneficios potenciales de la región están irremediablemente vinculados a su correcta extracción.

Como cualquier otra tarea de análisis, la segmentación en regiones es un proceso sujeto a inexactitudes en el procesamiento. Los errores producidos en las etapas tempranas del análisis

a nivel de píxel se corrigen usualmente en etapas posteriores. Por el contrario, no existe generalmente un proceso de marcha atrás para los errores cometidos durante la segmentación en regiones. Si, durante el proceso de segmentación, algunos de los contornos relevantes de la escena no se detectan, las entidades semánticas definidas por éstos son difícilmente recuperables con posterioridad. Esta es la principal razón para el diseño conservador aplicable a las soluciones propuestas. Por ejemplo, en el capítulo 4, la capacidad del método para detectar todos los contornos relevantes se priorizó sobre la precisión del método en detectar sólo los contornos relevantes. De manera similar, en el capítulo 7, las soluciones basadas en regiones definidas requerían de la existencia de un método complementario a nivel de píxel para guiar las decisiones. Con el mismo objetivo, en el capitulo 8 se propone analizar regiones a diferentes niveles de complejidad para incrementar la verosimilitud de obtener partes representativas del objeto en la fase de test. Finalmente, en el capítulo 9 se utilizó un esquema de segmentación difuso para evitar la asignación absoluta de píxeles a regiones, reduciendo los efectos de datos inciertos o fallos en el proceso de segmentación.

11.2 Resumen de los hitos y conclusiones principales.

El documento se ha estructurado en cinco partes (siendo esta la quinta). Analizaremos las conclusiones de cada parte por separado. Las conclusiones asociadas con los objetivos propuestos en la sección 1.2 del capítulo 1 se remarcan en negrita.

Parte I

La Parte I organizaba e introducía las aproximaciones en el documento haciendo especial énfasis en motivar el uso de la región para el análisis de imágenes y video.

Con este fin, en el capítulo 1 motivamos la región desde tres puntos de vista diferentes. Primero, establecimos la región como una unidad de análisis intermedia. Puesto que en la mayoría de las aproximaciones en el área de visión por computador, las resultados del análisis suelen agruparse para post-procesarse, cambiar el orden de operación parece una alternativa natural: primero agrupar para luego analizar. De este orden emergen dos espacios de descripción diferentes, el espacio de decisión (donde se agrupan los píxeles en regiones) y el espacio de características (donde se lleva a cabo el análisis). En segundo lugar, motivamos que el uso de la región puede derivar en una mejor de los resultados estadísticos de una aproximación. Finalmente, conectamos el procesamiento basado en regiones con las teorías de percepción humana en las que, aparentemente, se realizan procesos de agrupamiento de estímulos individuales para conducir el proceso de percepción.

El capítulo termina con la presentación de la organización del documento, sugiriendo órdenes de lectura alternativos y describiendo la evolución temporal de la investigación desarrollada.

Consideramos que esta descripción puede ayudar a comprender la diversidad de las ramas de investigación explorada a lo largo del documento.

El capítulo 2 está aún más orientado a la motivación del uso de regiones. El capítulo comienza con una **definición del concepto de región**. Con este fin, revisamos diferentes definiciones para la palabra *región* disponibles en varios diccionarios de inglés en línea. Fusionamos estas definiciones para obtener una definición genérica, pero cerrada del término región y del proceso de segmentación en regiones. A continuación, presentamos un conjunto de características básicas para describir una región en particular: su etiqueta, su representante, su contorno y la relación de adjacencia entre dos regiones. Consideramos que estas características son parte del proceso de segmentación en regiones porque resultan de éste. El capítulo sigue con una revisión de los retos específicos que han motivado el uso de regiones durante el desarrollo de esta Tesis. En particular, el uso de regiones está motivado por su habilidad para: reducir la brecha semántica, reducir el ruido, muestrear el espacio de características, suministrar entornos de descripción adaptables y, si son extraídas de manera difusa, codificar similitudes entre píxeles.

Aparte de su carácter motivador, en la Parte I también se presenta la terminología y los conceptos utilizados a lo largo del resto del documento. Además, en la Parte I se esboza nuestra interpretación de la región, hecho que hace que esta parte sirva como una estructura básica sobre la que se sustenta el resto del documento.

Parte II

La Parte II se centra en el proceso de segmentación en regiones (RS por sus siglas en inglés).

La Parte II empieza por el capítulo 3, en el que se revisan y **organizan las aproximaciones más relevantes en el ámbito de la segmentación en regiones**. El capítulo comienza con una discusión sobre la problemática asociada a la evaluación de las técnicas existentes para la segmentación en regiones. En particular, mostramos como diferentes observadores humanos perciben diferentes regiones cuando se les solicita segmentar una misma imagen. Esta diversidad implica que establecer un criterio común sobre la bondad de una segmentación es una tarea prácticamente irrealizable. El capítulo continua enumerando y describiendo los factores relevantes que identifican a una determinada aproximación a la segmentación en regiones, a saber: las características utilizadas para la descripción, la gestión de los bordes y de los contornos, el esquema de procesamiento y la escala a la cual se extraen las regiones. Posteriormente, se presenta la organización propuesta. En particular, proponemos organizar las aproximaciones existentes mediante un esquema dual. Primero, basándonos en las soluciones propuestas en cinco etapas (algunas de ellas obligatorias, otras opcionales): pre-procesado, extracción de características, análisis local, globalización y regionalización. Después, en función del nivel de procesado, distinguiendo entre enfoques locales, globales y combinados. Organizamos las técnicas existentes en base a esta última organización, mientras que las describimos siguiendo la organización por

etapas. Se establecen finalmente cinco grandes categorías. Tres de ellas conteniendo aproximaciones locales: clusterización, fusión de regiones y búsqueda de los modos. Consideramos que una categoría es suficiente para explicar las aproximaciones globales: minimización de energía. Finalmente, catalogamos las aproximaciones combinadas como basadas en grafos, una categoría que a su vez se subdivide en dos: jerárquicas y basadas en la detección de contornos. El capítulo finaliza con una descripción de los conjuntos de datos y métricas existentes para validar, en la medida de lo posible, la bondad de una determinada segmentación en regiones.

El capítulo 4 describe la integración de la técnica Mean-Shift (MS) con la teoría del espacio-escala. El capítulo comienza revisando ambos ámbitos en detalle, a fin de motivar el esquema propuesto y reforzar teóricamente las decisiones de diseño tomadas a lo largo del capítulo. En base a este estudio se propone **una estrategia para fijar automáticamente el ancho de banda espectral en un proceso MS**. Esta estrategia, diseñada para operar sobre distribuciones discretas de datos está inspirada en la teoría del espacio-escala y beneficia MS en tres aspectos: impide que el proceso se estanque en zonas planas de la distribución, acelera la convergencia de los procesos de búsqueda del modo y evita la convergencia a modos no globales. Adicionalmente, la estrategia diseñada elimina el proceso de selección del ancho de banda, parámetro que comúnmente determina la operación de los procesos MS. En su lugar establece un umbral relacionado con el mínimo número de muestras que deben ser asignadas a un modo para considerar éste como un modo global. El uso de la aproximación diseñada para la segmentación en regiones produce resultados dispares en función del set de datos analizados. Por un lado, en una base de datos compuesta por imágenes con texturas moderadas, el algoritmo diseñado es capaz de proporcionar regiones ajustadas a los contornos de los objetos en la escena respetando al mismo tiempo el detalle fino de la imagen. Comparativamente, la operación de la aproximación MS más utiliza en este escenario se muestra cualitativamente muy dependiente del valor del ancho de banda. Por otro lado, cuando utilizamos el algoritmo sobre un conjunto de datos compuesto por imágenes altamente texturadas, el esquema MS propuesto sigue respetando los contornos pero retorna severas sobre-segmentaciones de la imagen. Para paliar este problema, se propone un método jerárquico de fusión de regiones. Este proceso utiliza los representantes de color de las regiones en el espacio CIE-Lab para fusionar regiones bajo diferentes hipótesis de complejidad. Estas hipótesis se establecen mediante el análisis de la distribución de distancias cromáticas entre regiones, de manera que puede establecerse un conjunto genérico de hipótesis para imágenes de diferente naturaleza. Los resultados indican que el proceso de fusión de regiones, siendo beneficioso para el sistema en casos particulares, resulta en estadísticos muy por debajo de la mejor solución en el estado del arte. De estos experimentos, concluimos que el uso de información de color no es suficiente para afrontar el análisis de áreas altamente texturadas.

El capítulo 5 presenta un esquema para describir la variabilidad local alrededor de un píxel. El capítulo comienza con una revisión de los métodos existentes para describir la variabilidad

local, enfatizando en las métodos basados en Textones. Discutimos sobre las incongruencias teóricas de estos métodos a fin de presentar **la transformada discreta del coseno (DCT por sus siglas en inglés) como un banco de filtros fácilmente configurable (dependiente de un único parámetro, el tamaño del bloque) para describir la variabilidad local**. Por medio del tamaño de bloque, la naturaleza, la escala y el número de filtros totales quedan automáticamente definidos. Además, la DCT tiene la capacidad de aglutinar la mayoría de la información en las respuestas (coeficientes) de un pequeño subconjunto de filtros. Proponemos aprovechar esta capacidad y proponer un método para seleccionar automáticamente este subconjunto de respuestas relevantes para cada píxel analizando la representatividad de las respuestas. Analizando la sensibilidad de este proceso de selección en 200 imágenes llegamos a la interesante conclusión de que el número de filtros relevantes puede ser elegido igual para cualquiera de estas imágenes si las respuestas se ordenan por su relvancia y se toleran pequeños errores. Este estudio permite descartar un número sustancial de filtros, pero conlleva un problema si se requiere la comparación de las respuestas de los filtros seleccionados para cada píxel. Si bien el número de filtros es el mismo, la naturaleza de éstos filtros suele ser diferente para diferentes píxeles en la imagen, por lo tanto, la comparación directa de sus respuestas requiere de una comparación previa de los filtros. Para hacer frente a este problema, definimos una métrica para comparar cualesquiera dos filtros en el batnco de filtros de la DCT. La medida, aunque inspirada en premisas subjetivas, cumple con todas las condiciones para ser considerad una métrica. Utilizando esta métrica como base, derivamos una métrica mejorada que incluye además la respuesta de los filtros y considera el procesamiento multi-escala. La métrica así definida se utiliza para construir un mapa de contornos en el que se asigna a cada píxel una medida de su verosimilitud de ser parte de un contorno. Los resultados preliminares sugieren que el esquema puede ser útil para definir transiciones en la variabilidad local.

La enorme cantidad de aproximaciones a la RS disuade la creación de nuevas propuestas en el ámbito. Sin embargo, creemos que la mayoría de los problemas asociados al uso de regiones son consecuencia de errores cometidos durante su obtención. Por ello, enfocamos la segmentación en regiones desde una perspectiva minuciosa, y estudiamos los problemas de aproximaciones clásicas pero exitosas. En particular, en el capítulo 4 eliminamos la dependencia de MS de su parámetro más problemático. Esta eliminación se produce a expensas de la introducción de un nuevo parámetro que, en cualquier caso, consideramos más sencillo de establecer, aunque esta intuición debe aún ser demostrada al menos por métodos empíricos. Por otro lado, MS es una aproximación de bajo nivel; en nuestra esquema, convertimos MS en una aproximación de tipo combinado que utiliza información global para conducir el análisis local. En el capítulo 5 afrontamos un problema completamente diferente, en lugar de unir píxeles buscamos transiciones en la escena. Éstas se obtienen mediante el estudio de los cambios entre descripciones locales de los píxeles. En este ámbito, las aproximaciones más exitosas se basan en la aplicación de

un conjunto de filtros ingeniosamente diseñados, pero no consideran las relaciones entre los filtros para comparar sus respuestas. Las funciones cosenoidales sobre las que se genera la DCT permiten el establecimiento de relaciones directas entre los filtros que componen la transformada. Por ello, consideramos que la DCT es una herramienta simple sobre la que comenzar nuestro estudio sobre las relaciones entre filtros. En cualquier caso, como parte del trabajo futuro, está en nuestro ánimo la evaluación de sistemas de comparación similares sobre bancos de filtros alternativos. Adicionalmente, planificamos evaluar la solución propuesta en diferentes escenarios.

Parte III

La parte III aborda el uso de regiones como unidades de análisis complementarias para la sustracción de fondo (BS por sus siglas en inglés).

En el capítulo 6 **revisamos las aproximaciones recientes y relevantes en el ámbito de la sustracción de fondo**. Después de la definición del problema, el capítulo comienza con una descripción de los retos que debe afrontar una solución para BS. A continuación, organizamos las aproximaciones relevantes siguiendo un esquema basado en etapas, relacionando éstas con los retos antes descritos. Esta organización permite descubrir las fortalezas y debilidades de las aproximaciones existentes más allá de su evaluación cuantitativa. Además, se discuten los problemas de la evaluación cuantitativa, remarcando dos de ellos, asociados a la existencia de un único conjunto de evaluación apropiado: sobre-refinado de los resultados (que afecta a la forma del frente detectado para incrementar la precisión de los resultados) y diseño *ad hoc* de los métodos (que inhibe el estudio de los retos menos representados en el conjunto de evaluación). El capítulo finaliza con la descripción de las técnicas de evaluación existentes y con una discusión sobre la escasez de soluciones basadas en regiones en el estado del arte. En particular, concluimos que dicha escasez es debida al incremento en el coste computacional que implica el proceso de obtención de las regiones.

A pesar de la conclusión anterior, y bajo la expectativa de futuras mejoras tanto en el hardware como en el software de procesado, el capítulo 7 presenta **dos soluciones basadas en regiones cuyos resultados sugieren que un análisis a nivel de región puede beneficiar las soluciones de BS a nivel de píxel**. La primera de ellas se sustenta en un esquema MS muy sencillo para obtener regiones ciegas a la iluminación, a fin de que la influencia de los cambios de iluminación locales (sombras y reflejos) en la segmentación en regiones se reduzca sustancialmente. El esquema MS diseñado (germen del descrito en el capítulo 4) adapta el kernel de estimación a la búsqueda de continuidad del albedo y fusiona las regiones así obtenidas estudiando el alineamiento de sus vectores de color RGB. En global, el esquema es efectivo en la extrapolación de los resultados a nivel de píxel desde las áreas correctamente clasificadas as las áreas (conectadas) incorrectamente clasificadas. Por otro lado, la segunda solución amplia la primera considerando también la evolución temporal de las regiones. Con este fin, se propone

un esquema de modelado de fondo basado en regiones. El esquema depende de un modelado basado en matrices de covarianza de un conjunto configurable de características. Para detectar muestras de frente se propone un esquema de comparación basado en el uso de los autovalores de las matrices de covarianza. El modelo se define de naturaleza multi-capa para poder hacer frente a las variaciones temporales de las regiones de fondo. Los resultados preliminares sugieren que el método es capaz de suministrar máscaras de frente ajustadas a los contornos de los objetos sin necesidad de utilizar técnicas de post-procesado.

Estos esquemas se complementan con los apéndices A y B. El apéndice A describe un estudio de viabilidad sobre el uso de la caracterización basada en la DCT definida en el capítulo 5 y de su métrica asociada para incrementarla separabilidad entre frente y fondo. El apéndice B propone un modelado multi-capa multi-clase basado en un esquema de clases para inhibir la corrupción del modelo de fondo por muestras erróneamente clasificadas como frente.

La parte III y sus apéndices asociados presentan nuestras contribuciones a la sustracción de fondo. Si bien los resultados experimentales son prometedores, se requiere una evaluación más exhaustiva para alcanzar conclusiones sólidas. Sin embargo, el uso de regiones para la sustracción de fondo está obstaculizado por los requisitos de procesamiento en tiempo real que requieren la mayoría de las aplicaciones de BS. Por otro lado, creemos que existe una carencia de investigación en el procesamiento mutli-clase BS. Nuestro trabajo futuro buscará explotar el uso de información contextual para generar máscaras de relevancia / irrelevancia que podrían ser utilizados para definir un subconjunto de píxeles sobre los que realizar / evitar el procesamiento a nivel de región. En nuestra opinión, si la complejidad computacional de la segmentación en regiones se reduce mediante la disminución del número de píxeles a agrupar, el uso de aproximaciones basadas en regiones puede producir incrementos sustanciales en la calidad de los esquemas BS a nivel de píxel.

Parte IV

La parte IV versa sobre el uso de regiones para la descripción local (LD, por sus siglas en inglés) de puntos en imágenes.

El capítulo 8 comienza con una sección motivacional que busca establecer conexiones entre las teorías de percepción humana y las aproximaciones existentes en la visión por computador para la identificación de objetos. En esta sección motivamos una aproximación alternativa a las existentes para la identificación de objetos inspirada en la percepción humana. Con este fin, sugerimos que, a diferencia de la mayoría de las soluciones existentes en el estado del arte, la identificación de objetos no requiere de un amplio conjunto de datos de entrenamiento ni de un conjunto de modeles específicos por objeto. Por el contrario, hipotetizamos que el conocimiento puede generalizarse a partir de un conjunto pequeño de datos y que un único modelo puede ser usado para almacenar todo el conocimiento. Por un lado, la generalización del conocimiento se

alcanza particionando las instancias de entrenamiento bajo diferentes niveles de granularidad, **generando descripciones de las partes de un objeto, y suministrando así robustez a potenciales oclusiones hipotetizando en las partes visibles de un objeto**. Además para proporcionar también robustez a cambios en el punto de vista de captura, estas partes se alinean con las orientaciones de los puntos de interés que contienen. Para la descripción local de estas partes, proponemos dos estrategias de descripción novedosas que se sustentan en el enmascaramiento de descriptores bi y tri-dimensionales exitosos en el estado del arte. Por otro lado, el almacenamiento del conocimiento se realiza mediante codificación distribuida de las descriptores en un modelo neuronal y los objetos se identifican midiendo las respuestas a este modelo. Los resultados han sido extraídos asumiendo que los objetos han sido previamente segregados de una escena donde están severamente ocluidos, presentando, sin embargo, apariencias muy diferentes a las entrenadas. Considerando estas premisas, los resultados sugieren que el esquema propuesto is capaz de mejorar la operación del mejor esquema de LD en la tarea de identificación de objetos. Nuestro trabajo futuro se centrará en eliminar el pre-requisito de segregación.

El capítulo 9 propone una alternativa complementamente diferente para establecer correspondencias entre puntos entre imágenes. La solución **utiliza la calibración de la escena para restringir las posibles deformaciones geométricas entre imágenes de un entorno de descripción alrededor de un punto basado en regiones**. Las deformaciones del soporte son corregidas mediante el uso de homografías inducidas por planos. Las projecciones del soporte obtenidas mediante estas homografías constituyen los candidatos para la correspondencia. La bondad de estos candidatos se mide comparándolos con el entorno buscado bien por descripciones locales del color o del gradiente. Adicionalmente, se propone el uso de regiones difusas para definir la extensión del entorno de descripción y para pesar la influencia de las muestras dentro del entorno en consonancia con su similitud con el punto descrito. Además, y a fin de contemplar también transformaciones de la apariencia del soporte, estas regiones difusas se utilizan para pesar un esquema de transformación lineal de la descripción cromática del entorno de descripción. Los resultados, extraídos en escenarios de alta complejidad, son prometedores.

Los contenidos en esta parte están inspirados por la idea propuesta en Tola et al. [2010] sobre el uso de máscaras inhibidoras para hacer frente a oclusiones. El uso de regiones en lugar de patrones predefinidos permite adaptar el esquema de inhibición al contenido de la imagen. Cuando publicamos nuestra primera solución en esta línea (descrita en Navarro et al. [2014] y en el apéndice C), no eramos conscientes de que un esquema muy similar había sido diseñado (Trulls et al. [2013]). En el capítulo 8 aplicamos este esquema de enmascaramiento adaptativo sobre dos descriptores locales ampliamente utilizados. Mientras que, en el capítulo 9, utilizamos la idea propuesta en Trulls et al. [2013]. Este ámbito de inhibición de descripciones locales, es de hecho un campo candente de investigación actualmente, por lo que presente numerosas líneas

de investigación potenciales. Una parte sustancial de nuestro trabajo futuro estará enfocado en explorar algunas de estas líneas de investigación.

Part VI

Appendixes

Appendix A

A feasibility study of the use on the DCT for Background subtraction

The metric defined in Chapter 5 (included here in equation A.1) allows for a balanced comparison between two vectors containing different AC coefficients. This chapter aims to evaluate the feasibility of a background subtraction scheme driven by such metric. To this aim, we first define a background model in terms of AC coefficients and present learning scheme to capture the spatio-temporal variability of a pixel. We then evaluate the separation achieved by the model when classifying foreground and background samples. The proposed solution is compared with alternative features and alternative AC comparisons.

$$M\left[\psi_{x_1,y_1},\psi_{x_2,y_2}\right] = k_1\left[\mid x_1 - x_2 \mid \vee \mid y_1 - y_2 \mid\right] + k_2\left[\mid atan\left(\frac{x_1}{y_1}\right) - atan\left(\frac{x_2}{y_2}\right) \mid\right] \tag{A.1}$$

A.1 A background model exploiting local variability

Describing local variability

We aim to create a background model fed with a useful description of the local variability around every pixel, we first calculate the WxW DCT for each image pixel (a preliminary image padding based on reflections is performed to account for boundary pixels). Then, for every pixel at position (x, y), the resulting AC coefficients are ranked in descending order according to its energy, and the first N are selected. This results in a N length vector, $\overline{v}\,(x, y)$, each component containing two data: the AC coefficient value, $c\,(u, v)$, and the 2D index or identifier, (u, v), of the basis function to which it corresponds. Due to the complexity of establish a generic value of N for background subtraction scenarios, the N value is here defined as a user-settable parameter.

As N grows, the description sensitiveness increases, as low responses to the DCT basis functions are considered. However, the influence of noisy coefficients also increases with N, a situation that may bias the number of false classifications of background pixels as foreground. Then, the value of N should be chosen as a trade-off solution between sensitivity and accuracy.

Capturing the temporal evolution of local variability

The following step is to pixel-wise keep track or store the temporal evolution of a pixel's local variability. We store the statistics for each pixel identically and independently, which allows for an efficient and fast parallel scheme for the subsequent updating and discrimination phases. The proposed accumulator is intrinsically motivated by the fact that the range of the AC coefficients of a background area does not extremely vary in time, as it has been empirically evaluated in several works, including [Lamarre and Clark, 2002]. The whole process is here presented for a luminance image, its extension to colour images is leaved as part of the future work.

The data structure for a pixel (u_0, v_0) considers statistics for every component $n = 1..N$ of the vector that describes it, that is, for the higher energy AC coefficient, the second higher, and so on. Statistics include, for the n^{th} component: a 2D histogram, $H_n(x, y)$, to estimate the probability of every AC coefficient being the n^{th} higher energy one; and a 2D function, $C_n(x, y)$, which stores and updates, via a similarity-driven running average scheme, the value of every n^{th} higher energy coefficient taken into account in $H_n(x, y)$. Hence, for a new background pixel instance characterized by a vector $\mathbf{f}(u_0, v_0) = \{\mathbf{c}_{ranked}(u, v), \mathbf{\Psi}_{ranked}(u, v)\}$, being $f_n(u_0, v_0) = \{c(x, y), \psi_{x_0,y_0}\}$ its n^{th} vector component, its statistics are updated following:

$$H_n(x_0, y_0) = H_n(x_0, y_0) + 1 \tag{A.2}$$

$$C_n(x_0, y_0) = (1 - r)C_n(x_0, y_0) + (r)c(x_0, y_0) \tag{A.3}$$

where the updating factor, r, controls the influence of the new coefficient values by evaluating its similarity to those already stored and then, expected:

$$r = min\left(\frac{C_n(x_0, y_0)}{c(x_0, y_0)}, \frac{c(x_0, y_0)}{C_n(x_0, y_0)}\right) \tag{A.4}$$

The solution chosen to update r is inspired in the short-term video stability premise, i.e. changes in the sequence do not occur suddenly but gradually. Overall, this data structure conveys a pixel spatio-temporal description composed of N pairs, or *layers*, $\{\{H_n(x, y), C_n(x, y)\}, n = 1..N\}$ each storing the statistics of the n^{th} higher energy AC coefficient of an WxW block DCT around the pixel.

The spatio-temporal feature: measuring similarity to the captured evolution of local variability.

The direct use in a BS algorithm of the data structure described for each pixel in the previous Section is simply unmanageable. We here combine this data structure and the novel metric proposed in Chapter 5 to obtain a compact similarity measure between an incoming instance of a pixel's variability and its captured temporal evolution . This similarity measure will be the pixel feature that the later described BS algorithm is based on.

After computing for an incoming arbitrary pixel (u_0, v_0) the vector $\mathbf{f}(x_0, y_0)$, we proceed to evaluate the distance of the n^{th} component of such vector, $f_n(u_0, v_0) = \{c(x_0, y_0), \psi_{x_0,y_0}\}$ to the corresponding n^{th} stored pair or layer of the data structure, $\{H_n(x, y), C_n(x, y)\}$. Firstly, using equation A.1, the indicator:

$$d_n = \sum_{x=0}^{W} \sum_{y=0}^{W} H_n(x, y) M[\psi_{x,y}, \psi_{x_0,y_0}], (x, y) \neq (0, 0) \tag{A.5}$$

evaluates the distance of the ψ_{x_0,y_0} basis function to the 2D histogram, $H_n(x, y)$. Then, in order to also account for the coefficient values, we weight the previous indicator, hence obtaining a final distance to each layer:

$$D_n = \alpha_n d_n \tag{A.6}$$

where α_n evaluates, via the Jaccard distance, the relative relevance of the ψ_{x_0,y_0} basis function—i.e., the relative response to the basis function—in the n^{th} layer:

$$\alpha_n = 1 - \frac{C_n(x_0, y_0)}{\sum_{x=0}^{W} \sum_{y=0}^{W} C_n(x, y)}, (x, y) \neq (0, 0) \tag{A.7}$$

Up to this point, we have a compact measure, D_n, of the distance between a vector component and the corresponding layer of the data structure. We now proceed to evaluate the distance between the whole vector and its pixel captured evolution. We propose to weight each layer's distance by an estimation, β_n , of the relevance of the corresponding layer. One option is to directly give higher relevance to the first ranked layers, following a linear or any other fixed scheme. However, this would fail for DCT distributions in which the higher responses are very similar in energy, situation that might lead, due to noise, to different rankings among the most relevant AC coefficients. In order to tackle this situation, we propose to set the relevance according to the relative energy of the modelled layers, similar to the definition of α_n :

$$\beta_n = 1 - \frac{\sum_{x=0}^{W} \sum_{y=0}^{W} C_n(x, y)}{\sum_{\iota=1}^{N} \sum_{x=0}^{W} \sum_{y=0}^{W} C_n(x, y)}, (x, y) \neq (0, 0) \tag{A.8}$$

Once each layer relevance is estimated, the final feature, from now on *WRAC* (i.e., based on

intensity Weighted Ranked AC patterns), used to characterize each incoming instance of a pixel results:

$$WRAC\left(u_0, v_0\right) = \sum_{\iota=1}^{N} \beta_\iota D_\iota = \sum_{\iota=1}^{N} \beta_\iota \alpha_\iota d_\iota \tag{A.9}$$

This feature has been designed to be robust to small variations of the pixel's DCT distribution: the distance d_n evaluated in equation A.5 makes the feature robust to small perceptual variations in the N higher energy AC patterns; the scheme to capture coefficients intensity (equations. A.3, A.6 and A.7) acts smoothing the bin based distribution of the basis function storage; the effect of intensity outliers in the final feature computation is minimized by the relative weighting strategy defined at equation A.7; finally, the inter-layer relative relevance factor (equation A.8) diminishes the influence of alternative rankings of relevant AC coefficients in the final feature computation.

The presented feature does not account for background multi-modality. Some decision scheme is then needed to distinguish similarity changes produced by background pixel instances different than those expected, from changes produced by foreground pixel instances. This would be straightforward if the feature values followed any kind of common distribution when calculated frame-to-frame for background pixels of representative scenarios. The following Section presents a solution to model the distribution of the $WRAC$ feature, taking into account possible different modes, as well as an strategy to threshold its values based on its expected distribution.

From here in advance, the DCT window size has been set to $W = 8$, for efficiency reasons, and the sensitivity parameter has been set to $N = 7$, in order to demonstrate that, even by using so few AC coefficients, the proposed pixel characterization adequately allows for a stable background modelling and an accurate foreground discrimination.

Modelling the distribution of feature values

In order to automatically detect foreground pixel instances but avoiding the setting of a fixed threshold over the $WRAC$ feature defined in equation A.9, we describe the different values that this feature shows for background pixel instances by a pseudo-parametric density model.

Hypothetically, foreground pixel instances would result in low probability responses when compared to the background pixel's feature density model. This section is devoted to empirically demonstrate that the values distribution of the proposed feature for background pixel instances can be well estimated or modelled, and that it does not even follow a mono-modal distribution for static backgrounds, contrary to that assumed, for instance, for the luminance value of a static background pixel, which is suitable to be modelled by a Gaussian distribution [Wren et al., 1996]. Three different density estimation schemes have been tested: a Single Gaussian (SG), a Generalized Gaussian (GD) and a Mixture of Generalized Gaussians (MGD). Their ability

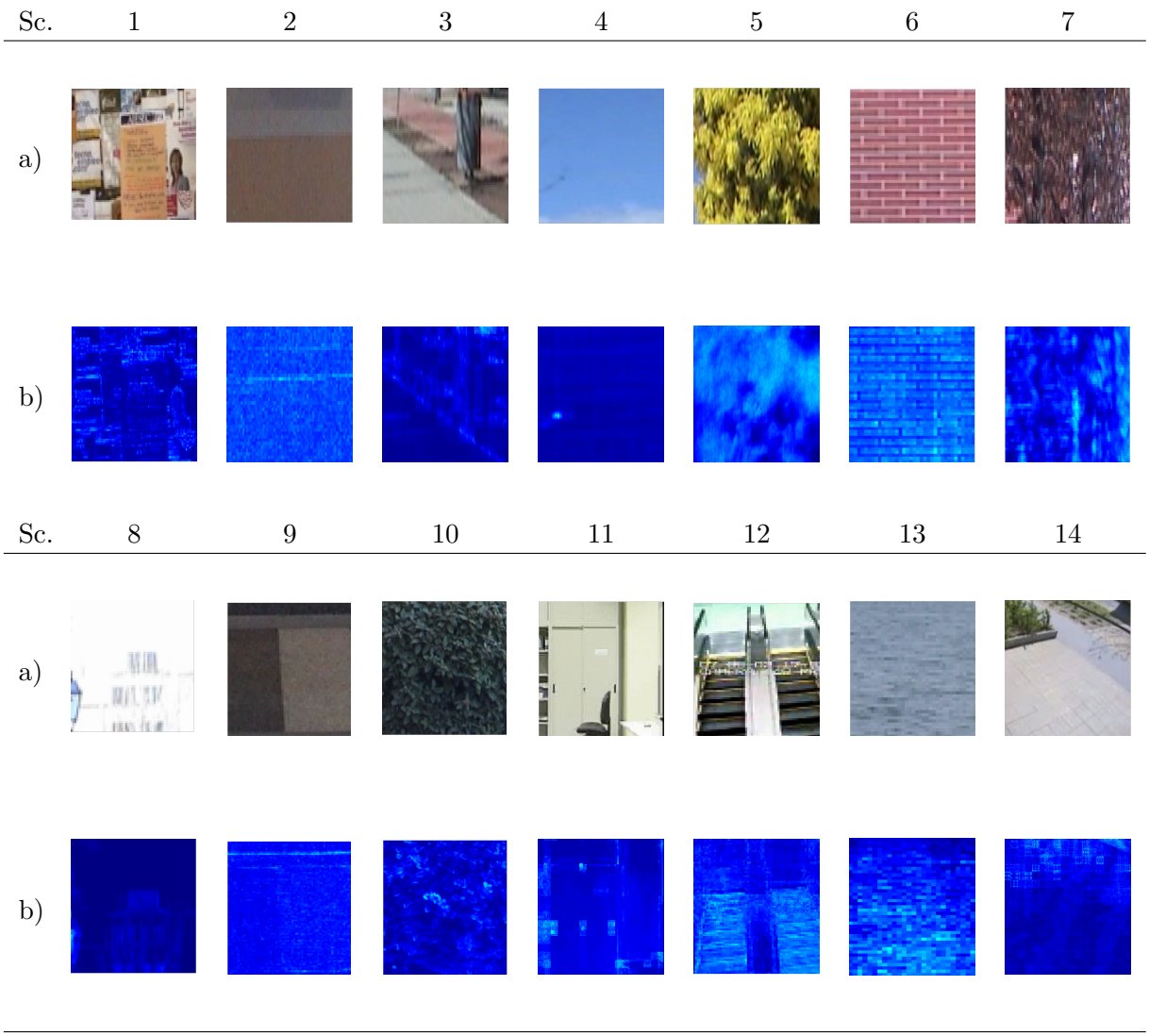


Fig. A.1. Background Scenarios to evaluate pixel variability. Row a) example frame. Row b) pixel variability.

to describe the distribution of the $WRAC$ feature has been measured via the Kullback–Leibler divergence between the estimated distribution and the real distribution, this last obtained from a representative ground-truth.

Obtaining the real distribution

Let $\overline{WRAC}$ be a vector containing the values of the the $WRAC$ feature for a given background pixel throughout a video sequence. In order to compute this vector only for background pixels, we use a data set, described below, with ground-truth segmented videos. For each video we calculate the normalized histogram of the $\overline{WRAC}$ vector using a bin size, BW.

Sc.	Related Complex Factor	Sc.	Related Complex Factor
1	Camera jitter, textured background	**8**	Illumination over-exposed area
2	Impulsive noise	**9**	Impulsive noise
3	Camera jitter, textured background	**10**	Multi-modal background
4	Homogeneous background	**11**	Moving Shadows, compression noise
5	Multi-modal background	**12**	Multi-modal background
6	Camera jitter, textured background	**13**	Multi-modal background
7	Multi-modal background	**14**	Moving Shadows, compression noise

Table A.1: Background complexity factors for cropped sequences (Sc: 1-14)

$$GD(\overline{WRAC}|\mu,\sigma,\beta) = \left[\frac{\beta\cdot\eta(\beta,\sigma)}{2\cdot\Gamma(1/\beta)}\right] exp(-\left[\eta(\beta,\sigma)\cdot|WRAC-\mu|^{\beta}\right]) \tag{A.10}$$

In order to be representative, our data set considers several background scenarios, including different complex factors. Scenarios are described in Figure A.1. Background variability in each scenario is illustrated including the average frame to frame square differences in Figure A.1 b): the brighter a pixel the higher its variability. Original full size videos, available at: [Tiburzi et al., 2008] S.1-S.10, [Prati et al., 2003] S.11, [Li et al., 2004] S.12, S.13 and [Benedek and Szirányi, 2007; Benedek and Szirányi, 2008] S.14, have been spatially cropped to select the video areas that show the complex factors we aim to solve, hence obtaining the respective smaller size sequences Sc.1-Sc.14. Background complexity of each cropped scenario is briefly described in Table A.1.

Fitting to the real distribution

As the expression that defines a Single Gaussian is generally known, we just include the expressions describing the Generalized Gaussian (GD) and the mixture of Generalized Gaussians distributions (MGD).

If $\overline{WRAC}$ follows a GD distribution with mean (μ), standard deviation (σ) and shape (β) parameters, then its density function is given by [Fan and Lin, 2009; Elguebaly and Bouguila, 2011; Krupinski and Purczynski, 2006] equation A.10, where

$$\eta(\beta,\sigma) = \frac{1}{\sigma}\left[\frac{\Gamma(3/\beta)}{\Gamma(1/\beta)}\right]^{1/2}$$

and $\Gamma(\cdot)$ is the Gamma function given by: $\Gamma(x) = \int_0^{\infty} t^{x-1}e^{-t}dt$, $x > 0$.

The shape parameter, β, controls the decay of the exponential and thus the shape of the distribution: the larger is β the flatter the distribution; while a small β describes a pointed distribution. Several methods to compute the shape parameter have been published; two of the

most popular are the Maximum Likelihood Estimator (MLE) [Varanasi and Aazhang, 1989] and the Moment Matching Estimator (MME) [Krupinski and Purczynski, 2006]. Even though [Fan and Lin, 2009] indicates that both methods have a poor resolution estimating high values of β , as the evolution of evaluated features is rarely uniform, we finally decided to use the [Krupinski and Purczynski, 2006] method, which is faster than [Varanasi and Aazhang, 1989].

In presence of a multi-modal background, $\overline{WRAC}$ might follow a MGD, which is a combination of K GDs, each weighed by a factor k_j, that can be defined as in [Elguebaly and Bouguila, 2011]:

$$MGD(\overline{WRAC}|\overline{\mu},\overline{\sigma},\overline{\beta},\overline{k}) = \sum_{j=1}^{K} k_j \cdot GD(\overline{WRAC}|\mu_j,\sigma_j,\beta_j) \tag{A.11}$$

In order to estimate the number of mixtures, K, and the weight of each mixture, k_j, for each pixel distribution, we propose to use a non-parametric kernel based density estimation technique, the Mean-Shift (a detailed explanation of the algorithm can be found in [Fukunaga and Hostetler, 1975] and [Comaniciu and Meer, 1999]). We use the Epanechnikov kernel (see chapter 4) and a bandwidth twice the bin size used in the histogram build (i.e., $2BW$) to minimize over-fitting. Mean-Shift parameters have been selected empirically, but remain equal for every test performed throughout this work. The main advantage of this approach is the automatic calculation of K.

Estimation results

As aforementioned, we use the Kullback–Leibler divergence (KL) to measure how precisely each estimated distribution fits to the real one. In general, the KL divergence between two discrete probability mass functions P and Q stands as:

$$KL(P,Q) = \sum_{j} P(j) \log\left(\frac{P(j)}{Q(j)}\right) \tag{A.12}$$

Table A.2 includes average values, for all the frame pixels of every sequence, of the KL divergence for every considered estimation distribution, which evaluate the suitability of each density estimation scheme for the task of feature modelling. KL has been computed just considering the frames with available ground-truth: sequences Sc.1-10 have available segmentation masks for every frame, while Sc.11-14 only have some of their frames segmented. In order to additionally compare the estimation of the proposed feature with that of a classical feature in BS techniques, a comparison with the fitting of the estimated density to the pixel luminance value is also included.

This analysis reveals that the distribution of the proposed feature is far from being Gaussian or even mono-modal, even when the pixel luminance value follows a Gaussian scheme (typical

Sc.	SG.		GD.		MGD.	
	Av.	Dev.	Av.	Dev.	Av.	Dev.
1	0.4183	0.0625	0.4236	0.0624	≈ 0	≈ 0
2	0.3795	0.0331	0.4002	0.0597	≈ 0	≈ 0
3	12.461	0.1236	12.455	0.1221	≈ 0	≈ 0
4	0.2967	0.0154	0.2967	0.0154	≈ 0	≈ 0
5	0.3829	0.0633	0.3842	0.0608	≈ 0	≈ 0
6	0.2173	0.1781	0.2173	0.1825	≈ 0	≈ 0
7	0.4962	0.1135	0.5102	0.1142	≈ 0	≈ 0
8	0.5005	0.0788	0.4682	0.0880	≈ 0	≈ 0
9	0.4313	0.0258	0.4325	0.0255	≈ 0	≈ 0
10	0.8826	0.1365	0.9115	0.1194	≈ 0	≈ 0
11	0.0815	0.1513	0.0806	0.1476	≈ 0	≈ 0
12	12.713	0.3399	12.449	0.3353	≈ 0	≈ 0
13	10.200	0.2540	10.093	0.2543	≈ 0	≈ 0
14	0.0870	0.1422	0.0869	0.1350	≈ 0	≈ 0

Sc.	SG.		GD.		MGD.	
	Av.	Dev.	Av.	Dev.	Av.	Dev.
1	0.9284	0.0911	0.7869	0.1165	0.0331	0.0072
2	0.8844	0.1321	0.6164	0.1395	0.0478	0.0150
3	0.8845	$\backsimeq 0$	0.6291	0.0709	0.0540	0.0062
4	0.8936	0.0701	0.6661	0.1031	0.0437	0.0150
5	0.8340	0.0795	0.6633	0.1125	0.0409	0.0041
6	15.461	0.2239	14.700	0.2486	0.0531	0.0095
7	0.9222	0.0702	0.7081	0.0848	0.0465	0.0080
8	0.9476	0.1555	0.6894	0.1569	0.0397	0.0292
9	0.8843	0.1422	0.5748	0.1624	0.0473	0.0096
10	0.9838	0.1471	0.7362	0.1572	0.0566	0.0116
11	0.6180	0.3875	0.5905	0.3739	0.1392	0.2274
12	14.256	0.3079	12.792	0.2903	0.1021	0.0250
13	17.196	0.0655	15.111	0.0919	0.1231	0.0195
14	0.4697	0.2175	0.4661	0.2232	0.0074	0.0102

Table A.2: *KL* divergence, average (Av) and deviation (Dev) between estimated and real distributions, for the proposed $WRAC$ feature (right) and for the pixel luminance (left).

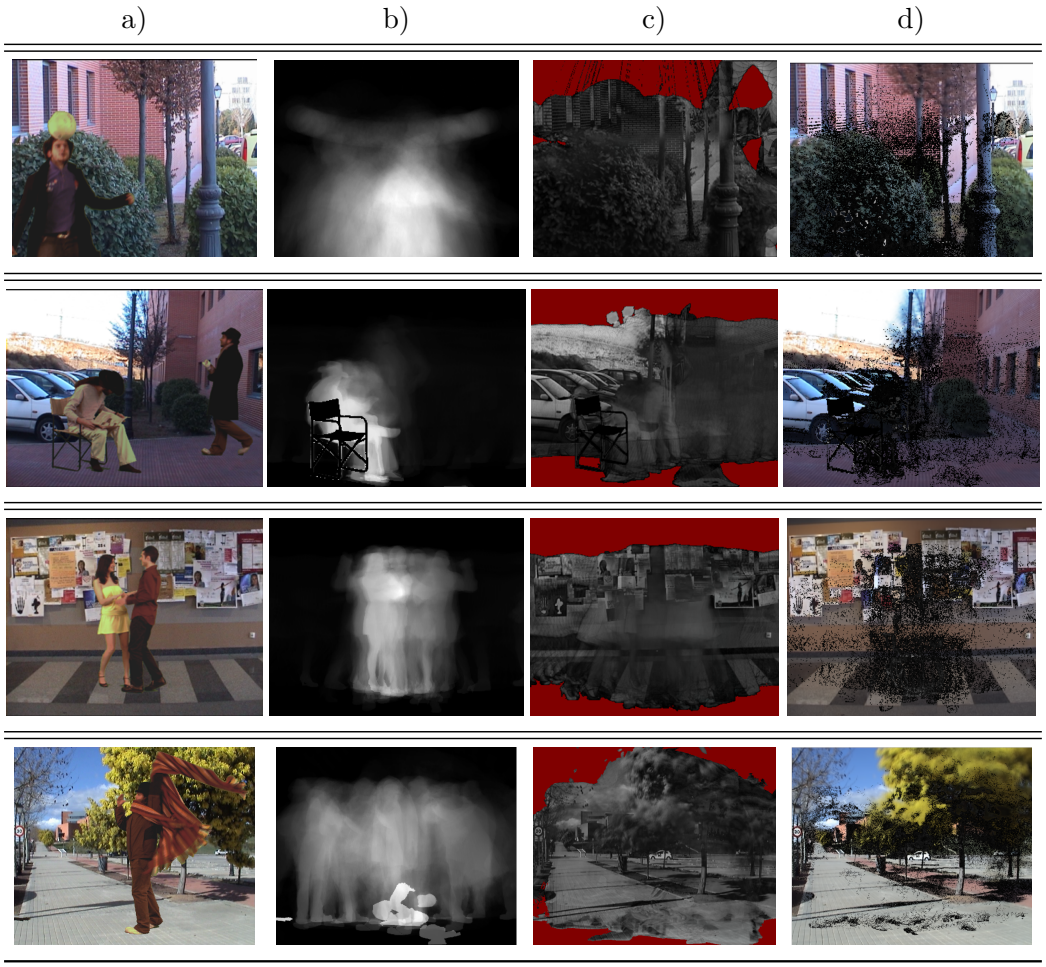


Fig. A.2. Example videos to measure foreground-background separability. Videos extracted from [Tiburzi et al., 2008]. a) Example frame, b) Foreground evolution, c) Average difference between foreground and background (red areas are never foreground), d) Prone to camouflage pixels (in black)

of static backgrounds). In this sense, results for Sc.11, where the background is mono-modal and only shadows and compression noise affect some of its pixels, luminance distribution is well described by a SG or a GD, while $WRAC$ can only be reasonably modelled with a MGD. Fitting results obtained for $WRAC$ over Sc. 11-14 are, due to the lack of a full descriptive set of annotated video samples, remarkably flimsier than those obtained for Sc. 1-10. Overall, the proposed MGD distribution estimation scheme perfectly models luminance distribution and also fits tight to the distribution of the proposed feature.

A.2 Separating Background and Foreground pixels

The previous Section showed the ability of the proposed feature to model the background. This Section includes several experiments to further demonstrate its ability to discriminate background pixels from foreground ones and its advantages respect to other similar features.

We have designed three different experiments targeting to evaluate the feature performance: raw feature separability, modelled feature separability and feature exportability. In order to fairly assess the potential discriminative power of the considered features independently of the common problems of the classical online model-learning BS techniques (i.e., hot starts, inaccuracies in the modelling, threshold selection, under and over modelling, etc.), feature models, when used, are updated using the ground truth segmentation; that is, model updating does not rely on any decision. The following experiments aim to empirically prove that the proposed feature presents a higher discriminative power and a higher robustness to common backgrounds changes than the other features evaluated.

For the first two experiments we have selected four videos from the data set described in [Tiburzi et al., 2008], where ground truth segmentation is available for every frame and contains several of the complex situations that affect backgrounds in real scenarios, which supports the robustness of the obtained results. These raw videos, described in Figure A.2, are 600 to 1200 frames long each, with 720x576 resolution. Apart from an example frame (a), we also include an average mask of foreground occurrence in the video (b), the average squared luminance difference between foreground and background for each pixel (uniform red areas correspond to frame areas not affected by the foreground) (c) and a frame showing the background pixels prone-to-camouflage (d). These refer to background pixels whose difference to the foreground is zero in at least one video frame, although larger differences might also cause camouflage. For the third experiment we use some more *popular* videos, but with ground truth segmentation just on some selected frames.

In order to compare the discriminative power of the $WRAC$ feature, four other features have been selected. Two of them aim to compare $WRAC$ against two alternative ways of considering DCT coefficients. One, which we will refer as $AC1$, replicates $WRAC$ but using the 2D Euclidean distance to measure the similarity between two DCT basis functions:

$$M'\left[\psi_{u_1,v_1}, \psi_{u_2,v_2}\right] = \sqrt{(u_1 - u_2)^2 + (v_1 - v_2)^2} \tag{A.13}$$

The other, which we will refer as $AC2$, replicates $AC1$ but using the first N coefficients of the DCT (following the classical zigzag order), instead of the N higher energy ones.

The third selected feature is the original uniform LBP Ojala et al. [2002], designed, as the proposed $WRAC$ feature, to measure local variability. Instead of the multilayer scheme there described, in order to evaluate the feature independently of a model, as aforementioned, we

	vs AC1			*vs AC2*			*vs LBP*			*vs Y*			*vs WRAC*			***vs all***	***Statistics***	
%	*w.*	*l.*	*t.*	*w.*	*l.*	*t.*	*w.*	*l.*	*t.*	*w.*	*l.*	*t.*	*w.*	*l.*	*t.*	***w.***	$\mu(t)$	$\sigma(t)$
AC1	-	-	-	*56.18*	*18.63*	*25.19*	*16.42*	*58.88*	*24.70*	*72.46*	*2.84*	*24.70*	*8.91*	*66.19*	*24.90*	***4.55***	*0.72*	*0.24*
AC2	*18.63*	*56.18*	*25.19*	-	-	-	*12.33*	*62.96*	*24.71*	*58.71*	*16.60*	*24.70*	*7.53*	*67.52*	*24.94*	***3.99***	*0.62*	*0.30*
LBP	*58.88*	*16.42*	*24.70*	*62.96*	*12.33*	*24.71*	-	-	-	*75.05*	*0.25*	*24.70*	*25.30*	*50.00*	*24.70*	***23.28***	*0.76*	*0.19*
Y	*2.84*	*72.46*	*24.70*	*16.60*	*58.71*	*24.70*	*0.25*	*75.05*	*24.70*	-	-	-	*0.11*	*75.20*	*24.70*	***0***	*0.43*	*0.37*
WRAC	*66.20*	*8.91*	*24.90*	*67.52*	*7.53*	*24.94*	*50.00*	*25.30*	*24.70*	*75.20*	*0.11*	*24.70*	-	-	-	***43.32***	*0.87*	*0.13*

Table A.3: Overall results for Foreground Background separability of raw data for proposed feature ($WRAC$), *Ranked Euclidean (AC1), ZigZag Euclidean (AC2)*, LBP and Luminance (Y) in terms of Bhattacharyya distance.

capture feature values for each background instance of a pixel in a histogram, which is then modelled . The circular radio around the pixel R_{region} that defines how many neighbours are used to build the LBP descriptor has been set to $W/2$ to perform a faithful comparison in terms of quantity of neighbours accounted. Finally, the fourth feature is the pixel luminance (Y), which has been, for years, the most popular way of considering the pixel value.

Raw feature separability

Experiment

This experiment evaluates the separability of the values that every feature yields for background and foreground pixels. We use the ground-truth segmented videos to obtain, for background pixel instances and for foreground ones , for every pixel position and for all the data set frames, the histograms or distributions of the values of the five features. Then the overlap for each feature between both distributions is evaluated using the well-known Bhattacharyya distance.

The comparison is performed in terms of wins (w), losses (l) and ties (t): given two Bhattacharyya distances, B_1 and B_2, resulting from computing the overlap between foreground and background distributions for features M_1 and M_2 respectively, M_1 beats (wins) M_2 if B_1 is higher than B_2, M_1 ties with M_2 if B_1 equals B_2, and losses if B_1 is lower than B_2. A Kolmogorov-Smirnov test with a 5% significance level is previously performed over each pair of background-foreground distributions in order to avoid comparison of identical distributions, a situation which finally did not occur in the selected data set. Comparisons between every pair of considered features, as well as overall winning and mean and standard deviations of pixel-average Bhattacharyya distances are included in Table A.3.

Discussion

Results indicate that the proposed $WRAC$ feature achieves higher Bhattacharyya distances than the others in more than a forty percent of the pixels evaluated. Taking into account that all features perform equal on pixel samples that are always background (the red areas depicted in

Figure A.2c), that sum-up to a 24.70% of the analysed pixels) , LBP performs better than the proposed feature in the 23.28 % of the cases, being the discriminative capability of the rest of the features much lower.

In average, observing the two far-right columns of Table A.3, the use of the proposed feature over just the seven highest ranked AC coefficients performs slightly better than the LBP descriptor applied over a $W/2$ pixels radius area around each pixel sample. It additionally outperforms the separability obtained with the raw luminance value of the pixel. Finally, the proposal to select the first N ranked AC coefficients and consider both their intensity value and the distance between the basis functions to which they correspond, results much more efficient than the two other options analysed via the $AC1$ and $AC2$ features.

While foreground-background discrimination is expected to be more effective as higher is the average Bhattacharyya distance and smaller its standard deviation, these results are not conclusive enough as medium average Bhattacharyya distances may also provide accurate foreground discrimination. The next experiment aims to overcome this observation.

Modelled feature separability

Experiment

This experiment simulates the behaviour of every feature in a BS technique, so that they can be compared in terms of recall and precision. We first use the distributions of foreground and background feature values obtained in the previous experiment for every pixel (u_0, v_0). We perform for each distribution the fitting tests described in Section A.1, in order to decide the density estimation model, among the proposed ones, which best fits to the distributions. As expected, the pseudo-parametric MGD results to be the most stable and best fitting procedure for all of them. Hence, this is the selected model.

We then start analysing each video sequence. Ground-truth data from the first 25 frames are used to train the MGD background model, $P_B(u, v)$, for each pixel. The foreground model, $P_F(u, v)$, which is hard to predict for a pixel, is initialized as a uniformly distributed function in the range of the possible values each feature can move in. Then, for every incoming frame we update the background or the foreground model for each pixel, based on the tag assigned to the pixel instance in the ground-truth. Finally, a specific pixel instance,(u_0, v_0), is tagged as background if $P_B(u_0, v_0) \geqslant P_F(u_0, v_0)$ and as foreground otherwise.Thus, models updating, which can be understood as an on-line learning procedure simulation driven by ground truth masks, is isolated from decision.

Individual results are given in Table A.4 for each feature in terms of Precision (Pre.) , Recall (Rec.) and F1-Score (FS). Additionally, percentage increase ($\triangle$) or decrease (when is negative) of the figures among the features are included.

	vs AC1			*vs AC2*			*vs LBP*			*vs Y*			*vs WRAC*			***Individuals***		
%	*△Pre.*	*△Rec.*	*△FS*	*△Pre.*	*△Rec.*	*△FS*	*△Pre.*	*△Rec.*	*△FS*	*△Pre.*	*△Rec.*	*△FS*	*△Pre.*	*△Rec.*	*△FS*	***Pre.***	***Rec.***	***FS***
AC1	-	-	-	*92.65*	*-38.66*	*90.95*	*-195.2*	*31.49*	*-78.14*	*7.23*	*1.94*	*6.22*	*-377.1*	*-3.3*	*-177.6*	***16.6***	***67.0***	***26.6***
AC2	*-1261*	*27.88*	*-1005*	-	-	-	*-3916*	*50.60*	*-1868*	*-1162*	*29.28*	*-936.0*	*-6392*	*25.51*	*-2967*	***1.22***	***92.9***	***2.40***
LBP	*66.12*	*-45.97*	*43.86*	*97.51*	*-102.4*	*94.92*	-	-	-	*68.57*	*-43.14*	*47.36*	*-61.63*	*-50.76*	*-55.83*	***49.0***	***45.9***	***47.4***
Y	*-7.792*	*-1.979*	*-6.64*	*92.08*	*-41.40*	*90.35*	*-218.2*	*30.14*	*-89.97*	-	-	-	*-414.3*	*-5.327*	*-196.1*	***15.4***	***65.7***	***25.1***
WRAC	*79.04*	*3.179*	*63.98*	*98.46*	*-34.45*	*96.74*	*38.13*	*33.67*	*35.83*	*80.56*	*5.058*	*66.22*	-	-	-	***79.2***	***69.2***	***73.9***

Table A.4: Overall results for foreground background separability of MGD based background models for; proposed metric ($WRAC$), 2D Euclidean (AC1), Zigzag Euclidean (AC2), LBP and Luminance (Y) in terms of Precision (Pre.), Recall (Rec.) and F1-Score (FS)

Discussion

In the light of these results, the proposed feature offers better figures than every other in terms of Precision, Recall (except for the inaccurate AC2) and F1-score averaged over the four analysed videos (see Figure A.2.). However, some other modelling or discrimination scheme may improve the results obtained for every feature. In this sense, in additional experiments, we have observed that variations in the sensibility parameter N do improve the foreground detection ratio of the proposed metric, but slightly affect the recall of the detection. This experiment only aims to show how, under the same conditions and using only the seven higher energy coefficients, $WRAC$ yield better results.

These results do not show the obtained segmentation masks nor the situations where the $WRAC$ feature behaves better or worse than the others; that is, a qualitative view of the results is still needed. Additionally, the set of videos used for this and the previous experiment offers a broad sample of background types and foreground appearances, but lacks in the variability of foreground relative size and specially, due to its chroma based nature, in the presence of illumination effects produced by the interaction of objects with the light sources. The following experiment tackles these concerns.

A.3 Feature exportability and qualitative results

Experiment

This experiment has been performed over some videos available in the public data sets described in [Li et al., 2004], in [Prati et al., 2003] and in [Benedek and Szirányi, 2007; Benedek and Szirányi, 2008]. We here compare the Y, LBP , and $WRAC$ features, all modelled via the Mean-Shift driven MGD. We formulate this experiment as a classification problem with two classes, background and foreground, where only the first class has been trained, using the annotated frames available for these videos

After training the models, each pixel sample is tagged with a probability indicating its

likelihood of belonging to the background model. No separation between training and test sets has been performed due to the small size of available data set and considering that background is occluded in most of the annotated frames. However, we believe that conditions are suitable for comparison as the process is the same for the three features.

Results are depicted in Figures A.3 and A.4, each including for a different sequence: a) An example frame, b) Receiver Operating Characteristic (ROC) curves, computed with statistics from all the available annotated frames, and selected points of work (PW) for each feature. c), d) and e) indicate the probability of each pixel (the brighter the higher) belonging to the background modelled using Y, LBP and $WRAC$ respectively. Finally, f), g) and h) show the segmentation masks resulting from threshold application the c), d) and e) probability mass functions at the PW, selected as that with minimum distance to the optimal behaviour (i.e., to the top-left corner).

Discussion

Figures A.3 and A.4 show that the proposed feature correctly separates between foreground and background instances of a pixel in presence of illuminations artefacts (i.e. reflection and shadows) and crowded environments (Highway, Sennon and Airport sequences), highly multi-modal backgrounds (Fountain sequence), strong umbra areas and camouflage situations (Highway sequence), strong illuminations changes (SwitchLigth sequence) and even in sequences severely affected by coding artefacts (Laboratory sequence).

As desired, the avoidance of the DC coefficient in the proposed feature enhances its robustness to local and global illumination changes. Additionally, the results indicate that the adequate behaviour of the proposed feature in camouflage situations (which produce holes in Y and LBP segmentation masks) does not affect the quality of the background modelling.

Finally, if we observe the results obtained in Figure A.4, a new camouflage situation in AC coefficient sense can be identified only for the proposed feature. This suggests that the proposed feature should ideally be combined with some complementary feature, for instance the photometric invariants of the dichromatic reflection model.

Computational efficiency versus the cut-off parameter, N.

Figure A.5 (right) includes average and maximum execution times per pixel (in milliseconds) as a function of N, and A.5(left) shows the average foreground-background Bhattacharya distance as a function of N, both over the Sennon sequence (see Figure A.4 left). Execution times have been measured on a C++ (OpenCV based) implementation running on a Intel Core 2 Duo (1.8 GHz) with 2GB RAM. We provide only per pixel time as the whole feature computation is parallelable (the operation over each pixel does not affect the others), so that it can be efficiently implemented (e.g. via a GPU based solution). According to the depicted curves, the feature

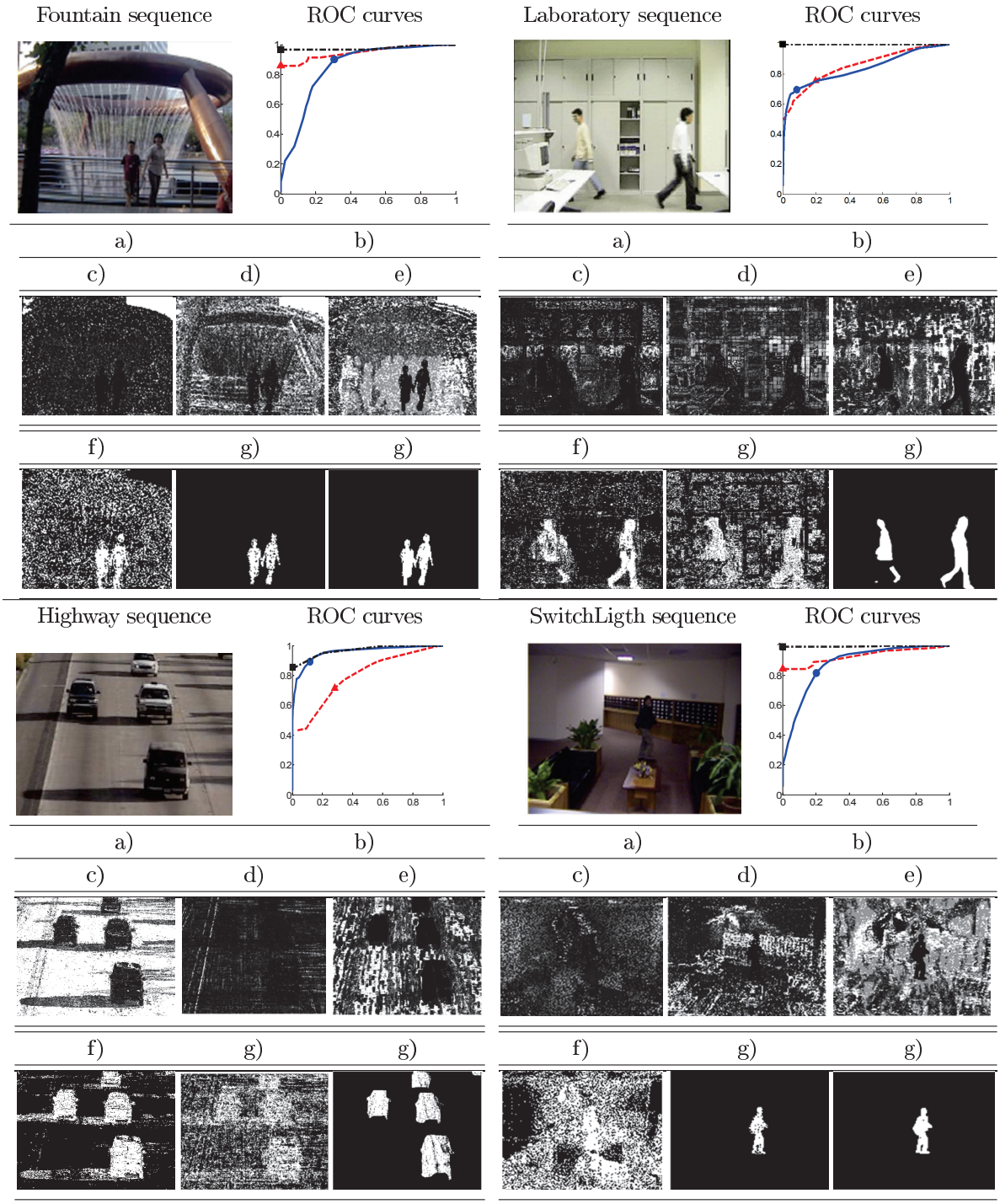


Fig. A.3. Qualitative Results 1: a) Example frame, b) ROC curves: LBP—solid blue line—WRAC—dashed-dot blue line—Y—dashed red line— c) Luminance distance d) LBP distance, e) WRAC distance, e) Threshold luminance at PW, f) Threshold LBP at PW, g)Threshold WRAC at PW

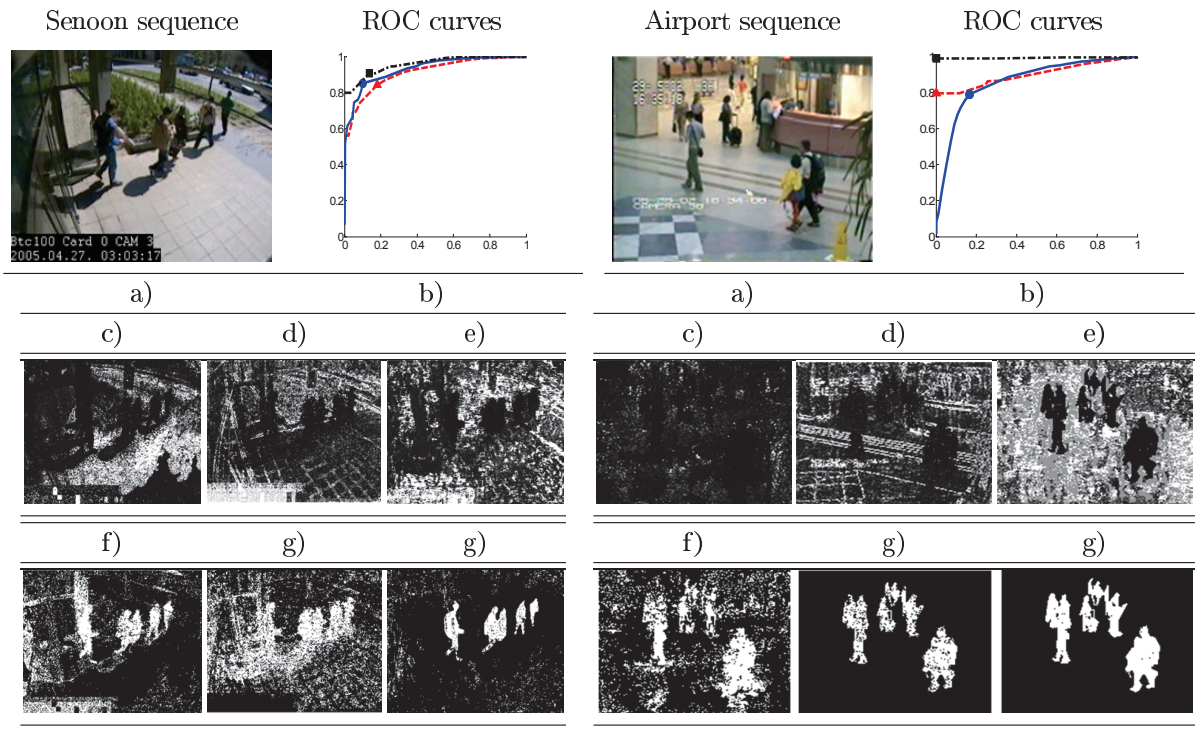


Fig. A.4. Qualitative Results 2: a) Example frame, b) ROC curves: LBP—solid blue line—WRAC—dashed-dot blue line—Y—dashed red line— c) Luminance distance d) LBP distance, e) WRAC distance, e) Threshold luminance at PW, f) Threshold LBP at PW, g)Threshold WRAC at PW

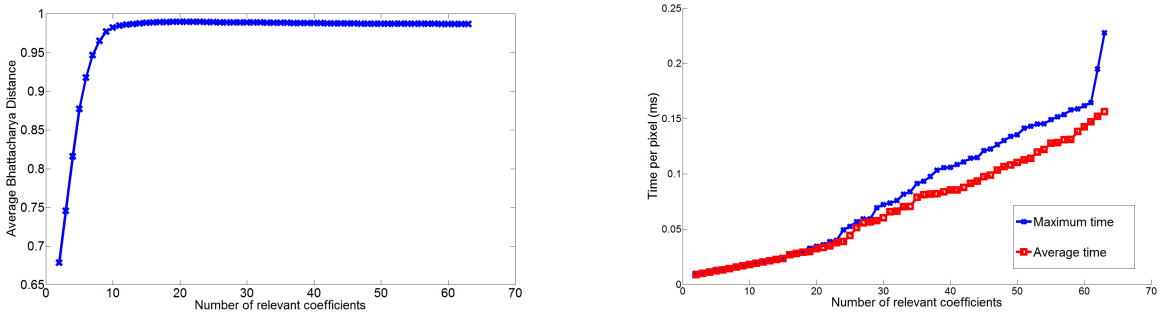


Fig. A.5. .Number of relevant coefficients vs Bhattacharya distance vs time per pixel

should be preferably designed in the elbow of the left figure (N between 5 and 11). In this area, the feature efficiently achieves high background-foreground discrimination.

A.4 Chapter conclusions.

This chapter presents a feature derived from the metric designed in chapter 5 to capture the evolution of local pixel variability, with the aim to help solving camouflage situations in BS techniques and, in general, to enhance foreground-background separability. Combined with an adequate background model, the feature has proven to operate successfully in the presence of illumination changes and multi-modal backgrounds.

The characterization of the pixel local variability is based on the selection of a variable set of DCT coefficients, those with higher energy. This requires the definition of a novel metric to evaluate the similarity between any two DCT basis functions, which, based on perceptual observations, equally balances variability rhythm and direction. The evolution of local variability is then modelled via a novel layer-based scheme. Finally, the comparison of every incoming instance of pixel local variability to the evolution model results in the proposed spatio-temporal feature, $WRAC$, with same dimensionality as the input data: a feature value per pixel.

Some of the advantages of the proposed feature derive from the way of capturing local variability:

1. The feature accounts for neighbouring pixels correlation, as DCT works at region level (i.e., a block) for each pixel.
2. The underlying AC coefficients that characterize the dominant variability are low correlated, due to the nature of the DCT.
3. It well balances rhythm and directional changes among AC coefficients, due to the proposed metric.
4. It is capable of handling medium-intensity illumination changes, as block size is usually smaller than changing areas.
5. Due to its region-based nature, it is more robust to clutter and occlusions.

Some other advantages, these derived of the scheme to handle variability evolution, include.

1. Feature sensibility is configurable by tuning the cut-off parameter, N.
2. Its compactness, a feature value per pixel, does not rely on the value of N.
3. The propose scheme favours temporally stable AC coefficients independently of their intensity while diminishes the influence of noisy ACs which evolution does not follow a stable pattern along time.

Feature capabilities are demonstrated via three different comparative experiments, which empirically evaluate the proposed feature, either standalone or included in a similar to the SoA background model, against common features used to describe pixel luminance and pixel local variability, and show robustness to camouflage and illumination changes. Results are presented for challenging public video sequences. Future work includes the definition, designing and testing of a scheme to combine $WRAC$ with complementary features.

Appendix B

Multi-class background subtraction

In this appendix, we propose a background subtraction video segmentation algorithm that works by modelling the different appearances of a pixel in a set of independent layers. The main contribution of this work with respect to the existing approaches is the use of an a priori classification scheme that classifies the pixel before updating the background model. This scheme isolates the pixel instances that belong to the foreground, hence avoiding their influence in the model updating and discrimination processes of the subsequent frames. The presented results demonstrate the successful performance of the algorithm in the presence of highly dynamic backgrounds, foreground–background similarity, hot starts and abrupt illumination changes. This work has been done in collaboration with Alfonso Colmenarejo.

B.1 Problem statement.

There are several strategies to model dynamic backgrounds in BS algorithms. The most popular is to describe the pixel evolution by a parametric model resulting in a combination of simpler sub-models (as the Gaussians in a mixture of Gaussians). These strategies are capable of handling several modes in a pixel value, one per sub-model Stauffer and Grimson [1999]. However, the updating of each sub-model affects the others. Alternatively, some authors propose the representation of the background model in k layers. For instance, similar to that proposed in Brault and Mohammad-Djafari [2004], for a frame at instant t, the likelihood of a new pixel sample, x_t, belonging to a layer, z_t, of the background, BG_t, may be given as:

$$p(BG_t, z_t|x_t) = p(BG_t, |z_t, x_t)p(z_t|x_t) \tag{B.1}$$

The main advantages of using multilayer schemes are: (i) modifications of the intra-layer models do not affect the rest of the layers, and (ii) the likelihood of a sample belonging to a layer, $p(z_t|x_t)$, and of a layer belonging to the background, $p(BG_t, z_t|x_t)$, are isolated. State-of-

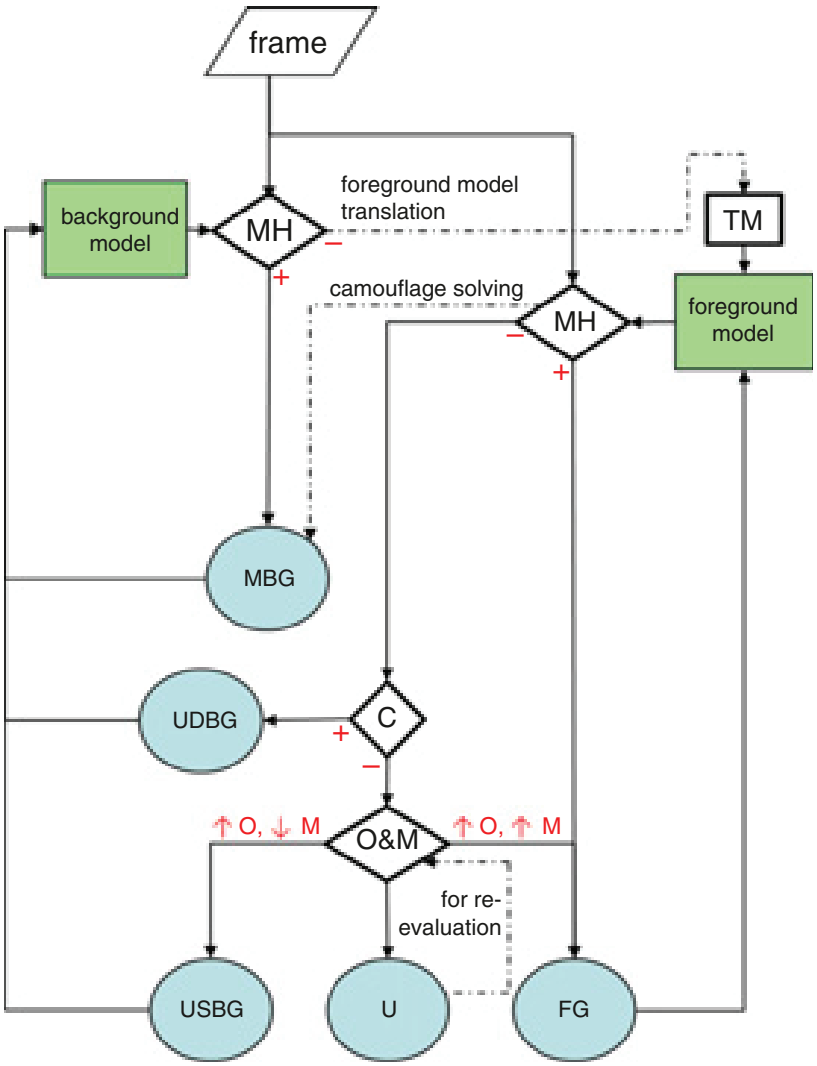
the-art algorithms in the task of multilayer background modelling use Bayesian-based schemes to update the model. Two different approaches can be differentiated. One is to have several layers modelling one class (background, shadow, foreground etc.), as in Porikli and Tuzel [2005], where class assignment is performed by thresholding the confidence of each layer being the modelled class. In this strategy, the thresholding stage is crucial as misclassification effects would be propagated in the model. An alternative approach is to model each class in one layer Benedek and Szirányi [2008]; here the matching process between each new sample and the layer is the key factor, as it determines the class. In this chapter, we present an hybrid strategy that combines both approaches, avoids the propagation of the effect of wrongly classified pixels in the model and increases the evidence for the matching process.

B.2 Pixel-based classification

A video pixel undergoes a series of pixel values or samples which should each be ideally assigned to the different identified classes.We define five possible classes for a pixel: modelled background (MBG), unmodelled dynamic background (UDBG), unmodelled static background (USBG), foreground (FG) and unclassified (U). MBG samples are those that fit in the established background model. UDBG samples are those that do not fit in the background model, but its previous and posterior samples are classified as MBG. USBG samples present still unmodelled appearance followed by equal appearance in subsequent samples instead of by samples classified as MBG. FG samples are those that fit in the established foreground model, and also those that do not fit in the background model and are followed by unequal appearance. Finally, U samples are potential samples of every other class; we store them in an intermediate layer for further analysis.

B.3 Classification procedure

The classification procedure is shown in Figure B.1 and described in the following Sections. Every pixel sample of an incoming frame has to be classified in one of the five considered classes (circles in Figure B.1): first, it is tested against the background and the foreground models to declare it either an MBG or an FG sample. If no match is found, the sample undergoes a set of on-purpose designed tests until a class is assigned. Class assignments are then used to update the models.



MBG	samples that belong to multilayer background model
UDBG	new samples showing oscillating appearance, indicating multimodal background that should update the background model
USBG	new samples with static appearance, indicating uncovered backgrounds, illumination changes etc. that should update the background model
FG	samples from moving objects that belong to foreground model
U	unclassified samples; need to track its evolution to further classify them in one of other four classes

Fig. B.1. Flowchart of the multi-class pixel-based background subtraction and pixel-class description. See text for details.

B.4 Background model

We propose a multilayer scheme inspired by Porikli and Tuzel [2005], where each layer models a pixel mode, so that background multi-modality is considered. A confidence measure is assigned to each pixel in each layer in order to evaluate its likelihood of being background. Confidence is proportional to the number of samples that fit in each pixel layer, and inversely proportional to the value of dispersion of such samples. As opposed to Porikli and Tuzel [2005], this confidence is not used to distinguish between reliable and unreliable background. Instead, layer confidence is used to evaluate the temporal confidence evolution of a pixel. In the proposed updating scheme, as foreground samples do not corrupt the background, every background layer is considered reliable, so that for MBG samples: $p(BG_t, |z_t, x_t) = 1$.

To classify new samples, all background layers are first ordered according to their confidence values. Intra-layer sample matching, $p(z_t|x_t)$, is performed by a full covariance Mahalanobis test (left MH in Figure B.1) or score. To avoid empirical thresholding of the score and to adapt to the appearance characteristics, we model the score evolution by a single Gaussian at each layer: a match is declared when a sample falls inside the Gaussian. These Gaussians are updated by a running-average scheme, with an envelope shape (i.e. the updating factor that weights the influence of new samples) that varies with the layer confidence. Low confidences indicate that the sample appearance is still being modelled, so that the Gaussian does not reliably represent the score evolution; then, a low factor is used to minimise the influence of outliers. As the confidence increases, the updating factor does the same in order to adapt the model to the progressive variation of the appearances. Finally, to avoid over-training, when a layer has reached a high confidence, the updating factor returns to the initial value and starts growing again.

B.5 Foreground model

The foreground model currently consists of just one layer. Again, each pixel in the layer has its own confidence value. For every incoming frame, the layer is translational-motion compensated (TM in Figure B.1). Motion parameters are estimated by comparing, via a simple but efficient Kalman filter, the stored foreground model mask with that resulting from all the frame pixels that do not match the background model. Once compensated, each incoming sample is tested (right MH in Figure B.1) against the foreground model. Matches are classified as FG even if they had been also classified as MBG; this achieves adequate reclassification of camouflaged pixels, especially in the presence of homogeneous foregrounds. Finally, FG pixel confidences are used so that, if the confidence value of a pixel decreases continuously, the pixel is removed from the foreground model.

B.6 UDBG detection

After background and foreground model comparison, the confidence evolution of the pixels associated with the remaining samples is evaluated (C in Figure B.1). Oscillations in an MBG pixel confidence value, observed in a temporal analysis window, indicate alternative periods of incoming samples matching and unmatching the model, which is the typical behaviour of a multi-modal pixel. Hence, the sample is classified as UDBG and is then used to initialise a new appearance or layer in the background model.

B.7 USBG and FG discrimination

Remaining samples may be USBG, FG or U samples. To discriminate among them two blob-based descriptors are computed. Blobs are extracted from the set of remaining samples (B_t blobs) and from the U or intermediate layer (B_{t-} blobs) and matched via a Kalman filter. If (x, y) are the co-ordinates of a blob's mass centre, the descriptors to base a decision (O&M in Figure B.1) are blob overlapping (O) and motion (M):

$$O = \frac{B_t \cap B_{t-}}{B_t \cup B_{t-}} \quad M = \sqrt{(x_{t-} - x_t)^2 + (y_{t-} - y_t)^2} \tag{B.2}$$

Samples that belong to a blob showing $O = 1$ and a $M \approx 0$ are declared USBG. New FG samples are declared for blobs showing $O > 0.5$ and $M > T$ where T has been empirically selected to be the 5% of the frame diagonal. The remaining samples are kept as U samples in the intermediate layer and are used for next frame classification.

B.8 Experimental results

Final segmentation masks include for each frame FG and U pixels. Results are obtained via evaluation of the dataset described in Tiburzi et al. [2008]. Videos are selected owing to their highly dynamic background, repetitive foreground and absence of shadows (as the presented approach does not cope with them). Metrics used for comparison are FScore1 (FS1) and FScore0 (FS0) as described in Herrero and Bescós [2009]. Quantitative comparison (Figure B.2) is performed against classical state-of-the-art methods (three single-layer models supporting multi-modality Stauffer and Grimson [1999]; Elgammal et al. [2002]; **?**, and a multilayer one using a Bayesian framework Porikli and Tuzel [2005]). Qualitative results of the system performance illustrating frequent background problems are included in B.3.

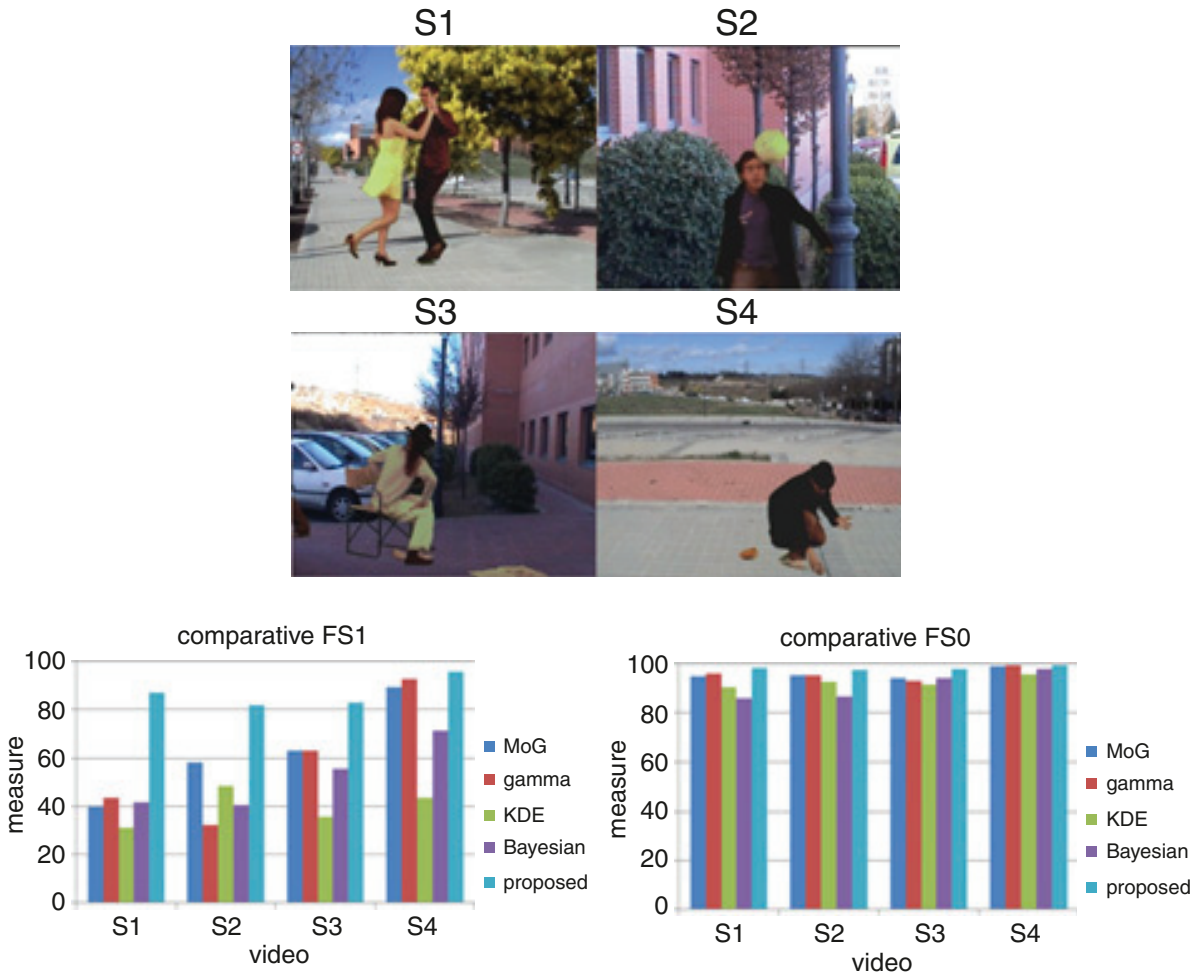


Fig. B.2. Comparative quantitative results

Fig. B.3. Comparative qualitative results

B.9 Discussion and future work

Results indicate that the proposed method achieves good performance in the presence of illumination changes, hot starts, camouflage situations and highly dynamic backgrounds. Additionally, it does not propagate miss-classified pixels and avoids the introduction of foreground in the model (as opposed to Porikli and Tuzel [2005]), which results in notably higher FS1 values. Further research must include the introduction of a cast shadows devoted layer.

B.10 Chapter conclusions.

Results indicate that the proposed method achieves good performance in the presence of illumination changes, hot starts, camouflage situations and highly dynamic backgrounds. Additionally, it does not propagate miss-classified pixels and avoids the introduction of foreground in the model (as opposed to Porikli and Tuzel [2005]), which results in notably higher FS1 values. Further research must include the introduction of a cast shadows devoted layer.

Appendix C

Super-pixel based isolation of the Scale Invariant Feature Transform

C.1 Introduction

Existing point-of-interest (POI) descriptions are biased by the information surrounding the point. Whereas in self-contained images this information is useful for enhancing the repeatability of the description, its use is inadequate for the description of objects that might be surrounded by variable backgrounds. To tackle these situations, a new POI descriptor—super-pixel-based isolation of the scale invariant feature transform (SP-SIFT)—is proposed. The classical SIFT descriptor is modified by isolating the information of the flat areas that compose it. It is proposed to include super-pixel information in the description stage of the SIFT. The obtained results suggest that a so-built descriptor increases the repeatability of SIFT points in these scenarios while keeping its robustness to global transformations of the image: blurring, changes in viewpoint, scale and lighting. The method is presented here as an extension of the SIFT. However, the idea behind it may be easily exported to most of the existing POI-descriptors in the state-of-the-art. This work has been conducted together with Fulgencio Navarro.

C.2 Main idea and motivation

Despite the wide range of point-of-interest (POI) descriptors reported in the literature, the automatic characterisation and matching of POIs is still an unresolved issue. Owing to a combination of reasonable performance and publicly available implementations, the scale invariant feature transform (SIFT) Lowe [1999], speeded up robust features Bay et al. [2006] and DAISY Tola et al. [2010] are the most popular description techniques in this field. These three techniques share a similar relative-to-neighbourhood approach for their description stage. This approach,

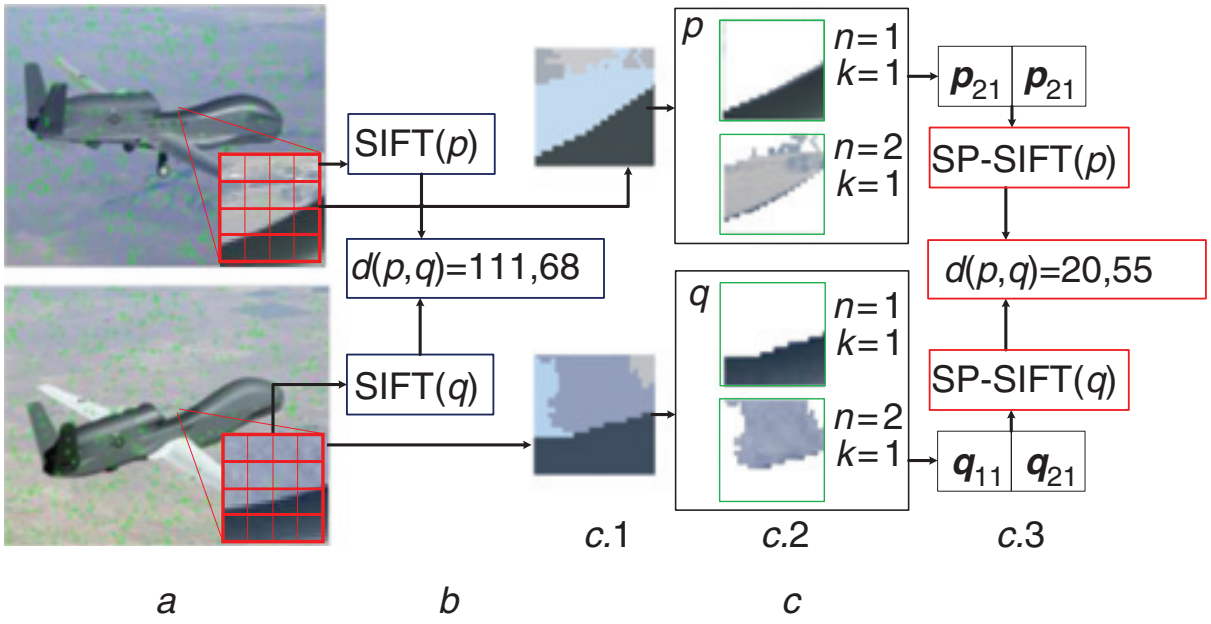


Fig. C.1. Graphical scheme of SP-SIFT operation.

while effective for the description of self-defined entities such as images or areas inside objects, gives inadequate results for the description of objects that might be surrounded by variable backgrounds (object collections, object tracking etc.). In such a case, the descriptor area of a POI in the object's boundary will not necessarily resemble that of a corresponding POI in another image.

This situation is partially illustrated in Figure C.1. The object of interest, the plane, is surrounded by different backgrounds. At image-scale, the SIFT is prone to detect several POIs whose descriptors include both the plane and the background. As the plane area is flat, these descriptors mainly capture the transition between the plane and the background and the texture of each background, a situation which can lead to a poor match. To eliminate the background information of the descriptor, we propose to include region information in the SIFT descriptor, which in the example achieves a five times better match.

C.3 Links with previous approaches

There are few studies that proposed to use region information to improve POI description methods. They can be roughly divided into two branches: algorithms that propose to describe regions with point descriptors and algorithms that include region information in the POI description stage. The systems in the former replace the POI detection stage with a region partition of the image, which diminishes the representativeness and stability properties derived from POI

searching Mikolajczyk et al. [2005]. Whereas, the approaches in the latter are commonly application oriented; for instance, in Tola et al. [2010], the DAISY description pattern is adapted by an occlusion detection method Boykov [2001] in order to avoid the inclusion of information from the occluding regions in the description of the occluded ones. The regions are also used as a post-processing technique to refine POI matching Suga et al. [2008].

Alternatively, POI methods are used as region segmentation assistants, as in Kudo et al. [2012]. However, our proposal lies far from all these studies as it includes the region information in the core of the POI description stage. Whereas, the problems described in all these articles support the motivation of our proposal, to our knowledge little research has explored this specific line.

C.4 SP-SIFT

As in the original SIFT method, the process is divided into two stages: detection and description. The detection stage follows the SIFT method (see Figure C.1 a). Then, the image is segmented into super-pixels (Figure C.1, c.1) by using the algorithm described in Achanta et al. [2012]. This region segmentation method was selected due to its superior conservation of the contour information of the scene Achanta et al. [2012]. It is important to remark that, due to the characteristics of both methods, the SIFT points are expected to be located in the super-pixel's boundaries.

The core of SP-SIFT lies in the description stage. Let us define the concept of the 'active' area as the surface of a super-pixel that overlaps with the SIFT square description area (see Figure C.1, c.1 and c.2). In SIFT, the descriptor and the principal orientation (used later for description normalisation) are both computed over the gradient information of the whole description area (see Figure C.1 a and b).

We propose to evaluate them separately for each active area (as in Figure C.1 c.2 and c.3). More in detail, for every active area in the SIFT square description area: (i) The pixels out of the active area are removed and the gradient information of this modified square description area is extracted. (ii) The principal orientation is computed. (iii) A SIFT descriptor is obtained and normalised for every principal orientation, if more than one.

To avoid the description of size-marginal areas, active areas smaller than a quarter of the SIFT description area are discarded. Experimentally, we have observed that such a threshold represents a trade-off solution between a description's repeatability and distinctiveness. This process results in a set of SIFT descriptors – one or several up to four active areas – per detected SIFT point, which overall conform to the proposed SP-SIFT descriptor.

Each of these descriptors describes an active area. We do not know a priori which active area describes a SIFT point in the best manner. Therefore, we defined the best area as the one

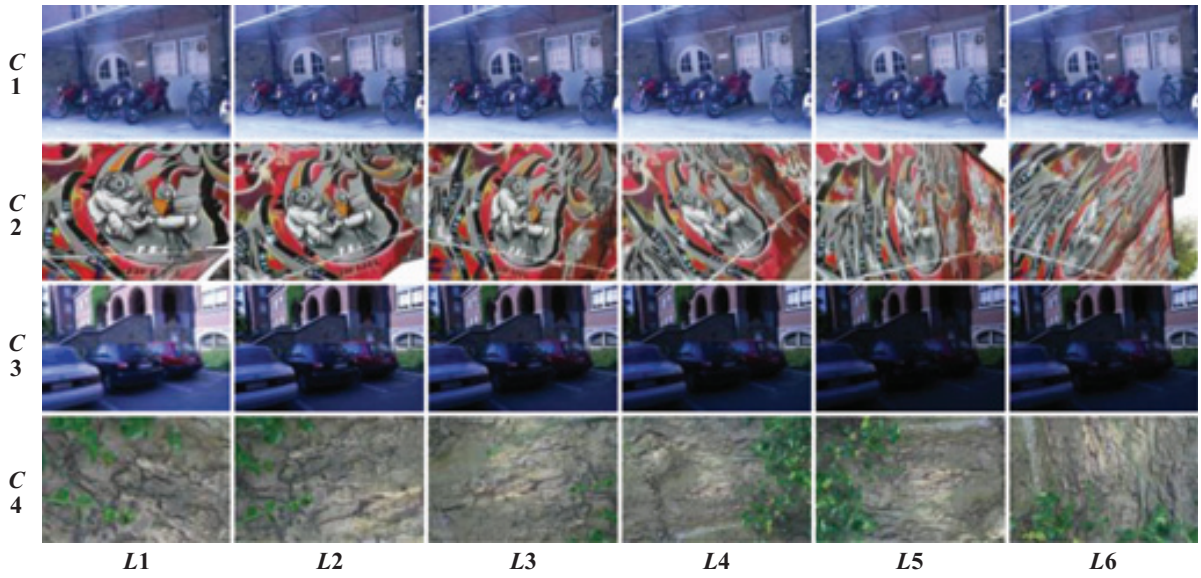


Fig. C.2. Image perturbation data-set.

that minimises the distance between the SP-SIFT descriptors.

To evaluate the matching of the SP-SIFT descriptors of two POIs, $\mathbf{p}$ and $\mathbf{q}$, we propose the following approach: let $\mathbf{p}_{n,k}$ be the k^{th} descriptor, $k = (1 \dots K)$, of the n^{th} active area, $n = (1 \dots N)$, of the point $\mathbf{p}$, where K depends on the number of the principal orientations and $N \leq 4$; similarly, let $\mathbf{q}_{n',k'}$ $(k\prime = (1 \dots K\prime)$, $n\prime = (1 \dots N\prime)$ and $(N\prime \leq 4)$ be the descriptor of the point $\mathbf{q}$. As the cardinals of the sets of descriptors for $\mathbf{p}$ and $\mathbf{q}$ might be different, we propose to evaluate the similarity between $\mathbf{p}$ and $\mathbf{q}$ as the minimum distance between their respective descriptors:

$$d(\mathbf{p}, \mathbf{q}) = \min_{n,n',k,k'} \left(\left\| \mathbf{p}_{n,k} - \mathbf{q}_{n',k'} \right\|_2 \right) \tag{C.1}$$

, where $\|\mathbf{x} - \mathbf{y}\|_2$ stands for the Euclidean distance between $\mathbf{x}$ and $\mathbf{y}$.

C.5 Experimental results

To evaluate the proposed algorithm, we present two different comparisons against the SIFT. First, and following Mikolajczyk and Schmid [2005], we compare the stability of the SIFT and the SP-SIFT descriptors against the image changes (stability test). Then, we test the foreground–background-segregation capability of the SP-SIFT descriptor in an object's description problem (segregation test).

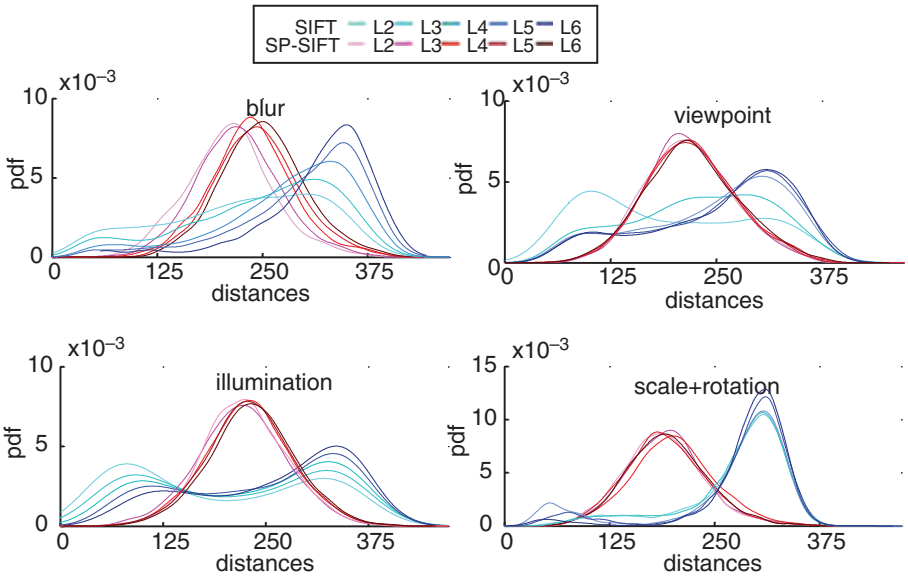


Fig. C.3. Stability test. Distribution of matching distance per category

Stability test

SIFT and SP-SIFT descriptors are extracted for SIFT points detected on the dataset presented in Mikolajczyk and Schmid [2005]. The dataset (see Figure C.2) agglutinates four different categories of image perturbations: image blurring ($C1$) and changes in the viewpoint ($C2$), illumination ($C3$), objects' scale and rotation (fused in category $C4$). Stability against compression changes is left out of the analysis as SIFT does not claim to be robust to this perturbation.

There are six images associated with each category (adding up 24), each increasing the complexity of the previous one (from the image $L1$: non-affected, to $L6$: most-affected). For a given category, the points in L1 are matched to their closest points in each other's image by using the distance proposed in Lowe [1999] (SIFT) or by using the proposed approach for SP-SIFT (defined in equation C.1), both following an injective scheme.

Figure C.3 shows an estimation of the density of the resulting matching distances, fit by a normal kernel with support constrained by the minimum and the maximum distances obtained for the category. The results in Figure C.3 suggest that the SIFT is not as stable as expected to the image perturbations: the density distributions shift to the right with complexity. On the other hand, the SP-SIFT seems to be more robust to these variations, as its distances distribution remains almost unaffected by complexity.

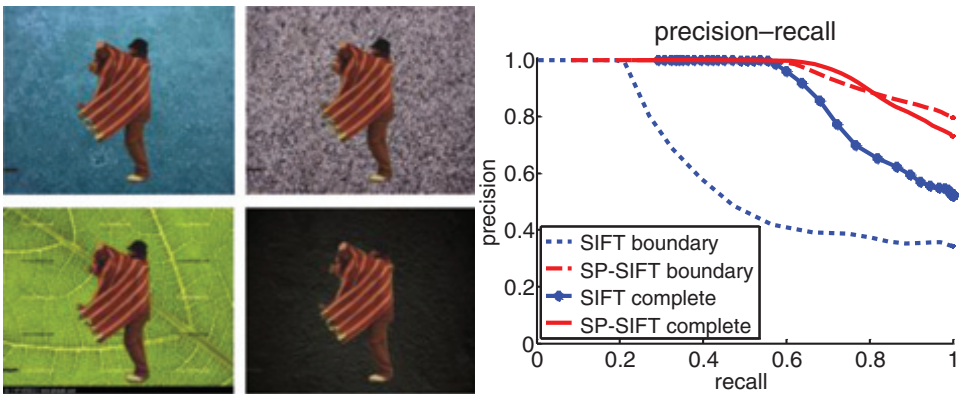


Fig. C.4. Segregation test.Left: analysed images. Right. Averaged Precision-Recall curves for objects and boundaries (solid lines) and exclusively for boundaries (dashed lines).

Segregation test

SIFT and SP-SIFT descriptors are extracted for the SIFT points detected on a toy example (see Figure C.4). A textured real object is superimposed over four highly textured backgrounds. Every so-built image is compared against each other, adding up to six comparisons. We again follow an injective point-to-point matching-scheme. The experiment is twofold. First, we evaluate the ability of each descriptor in matching all the object's POIs. Then, we focus only on the matching of the object's boundary POIs, which are affected by the foreground–background effects.

The results are illustrated in Figure C.4, via a classical precision–recall study by applying a threshold on the matching distance. In the light of these curves, the SP-SIFT outperforms the SIFT in both the experiments. In the task of the object's description, the SIFT generally yields to lower distances than the SP-SIFT when the POIs are fully contained inside the objects (reflected also in the lower modes in Figure C.3). However, the SP-SIFT discriminates these POIs in a better manner with respect to the background POIs, then ranking equally (or better) at the object's-inside POIs. The main differences between the SIFT and the SP-SIFT arise in the boundary points: the SIFT's description of these points hinders its overall operation for the reasons aforementioned, whereas the SP-SIFT adequately isolates the object's information. Although this toy example may not be representative enough, it reasonably shows the advantages of describing the object with the SP-SIFT descriptor, especially in the description of the boundary points.

C.6 Chapter conclusions.

This Chapter proposes the SP-SIFT to overcome the SIFT's limitations in scenarios where the description of the object of interest is disturbed by the surrounding information. This is achieved by the use of the tight-to-object super-pixels that drive the isolation of the object's parts in the description and allow its reorganisation. The benefits of the SP-SIFT in terms of description stability and discriminability are illustrated through two experiments. Essentially, this chapter proposes a new description algorithm that improves the operation of the SIFT in tasks that–in spite of being interspersed with the SIFT references–were out of its initial scope.

Glossary

RS *Region Segmentation*

RGB *Red Green and Blue (colour space)*

HSV *Hue Saturation and Value (colour space)*

BS *Background Subtraction*

MS *Mean Shift*

DCT *Discrete Cosine Transform*

CCD *Charge-Couple Device*

PCA *Principal Component Analysis*

MoG *Mixture of Gaussians*

KDE *Kernel Density Estimation*

LBP *Local Binary Patterns*

LBSP *Local Binary Similarity Patterns*

ROC *Receiver Operating Characteristics*

SIFT *Scale-Invariant Feature Transform*

SHOT *Signature of Histograms of OrienTations*

CGA *Colour Graphics Adapter (frame resolution)*

Bibliography

H. Aanaes, A. Dahl, and K. Steenstrup Pedersen. Interesting interest points. *International Journal of Computer Vision*, 97:18–35, 2012. [Cited on pages 202, 221, and 228.]

L. Abbott, E. T. Rolls, and M. J. Tovee. Representational capacity of face coding in monkeys. *Cerebral Cortex*, 6(3):498–505, 1996. [Cited on page 164.]

M. M. Abdelsamea, G. Gnecco, and M. M. Gaber. A survey of som-based active contour models for image segmentation. In *Advances in Self-Organizing Maps and Learning Vector Quantization*, pages 293–302. Springer, 2014. [Cited on pages 27 and 39.]

R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. SÃŒsstrunk. SLIC Superpixels Compared to State-of-the-art Superpixel Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274 – 2282, 2012. ISSN 0162-8828. doi: 10.1109/TPAMI.2012.120. [Cited on pages 42, 48, 49, 50, 57, and 301.]

J. Aloimonos, I. Weiss, and A. Bandyopadhyay. Active vision. *International Journal of Computer Vision*, 1(4):333–356, 1988. ISSN 0920-5691. doi: 10.1007/BF00133571. URL `http://dx.doi.org/10.1007/BF00133571`. [Cited on page 40.]

S. Alpert, M. Galun, A. Brandt, and R. Basri. Image segmentation by probabilistic bottom-up aggregation and cue integration. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(2):315–327, 2012. [Cited on pages 48, 53, 55, 57, and 81.]

A. Anand, H. S. Koppula, T. Joachims, and A. Saxena. Contextually guided semantic labeling and search for three-dimensional point clouds. *The International Journal of Robotics Research*, page 0278364912461538, 2012. [Cited on pages 251 and 261.]

P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. From contours to regions: An empirical evaluation. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2294–2301. IEEE, 2009. [Cited on pages 42, 44, 99, and 100.]

P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour detection and hierarchical image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(5):898–916, 2011. ISSN 0162-8828. [Cited on pages 41, 42, 43, 44, 45, 48, 53, 54, 55, 57, 78, 81, 84, 85, 87, 98, 112, 119, 166, 171, and 188.]

P. Arbelaez, B. Hariharan, C. Gu, S. Gupta, L. Bourdev, and J. Malik. Semantic segmentation using regions and parts. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3378 –3385, june 2012. [Cited on pages 161 and 193.]

D. Baltieri, R. Vezzani, and R. Cucchiara. Fast background initialization with recursive hadamard transform. In *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*, pages 165–171. IEEE, 2010. [Cited on page 131.]

K. Barnard, V. Cardei, and B. Funt. A comparison of computational color constancy algorithms. i: Methodology and experiments with synthesized data. *Image Processing, IEEE Transactions on*, 11(9): 972–984, 2002. [Cited on page 42.]

C. Barnes, E. Shechtman, D. Goldman, and A. Finkelstein. The generalized patchmatch correspondence algorithm. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *Computer Vision ECCV 2010*, volume 6313 of *Lecture Notes in Computer Science*, pages 29–43. Springer Berlin Heidelberg, 2010. ISBN 978-3-642-15557-4. doi: 10.1007/978-3-642-15558-1_3. [Cited on pages 206 and 207.]

O. Barnich and M. Van Droogenbroeck. Vibe: A universal background subtraction algorithm for video sequences. *Image Processing, IEEE Transactions on*, 20(6):1709–1724, 2011. [Cited on pages 132 and 133.]

H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *Computer vision–ECCV 2006*, pages 404–417. Springer, 2006. [Cited on pages 166, 206, and 299.]

C. Benedek and T. Szirányi. Study on color space selection for detecting cast shadows in video surveillance: Articles. *Int. J. Imaging Syst. Technol.*, 17:190–201, October 2007. ISSN 0899-9457. [Cited on pages 278 and 285.]

C. Benedek and T. Szirányi. Bayesian foreground and shadow detection in uncertain frame rate surveillance videos. *IEEE Transactions on Image Processing*, 17(4):608–621, 2008. [Cited on pages 96, 278, 285, and 292.]

H. Bhaskar, L. Mihaylova, and A. Achim. Video foreground detection based on symmetric alpha-stable mixture models. *Circuits and Systems for Video Technology, IEEE Transactions on*, 20(8):1133–1138, 2010. [Cited on page 134.]

I. Biederman. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94:115–147, 1987. [Cited on pages 160 and 161.]

L. Bo, K. Lai, X. Ren, and D. Fox. Object recognition with hierarchical kernel descriptors. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1729 –1736, june 2011. [Cited on pages 161 and 162.]

T. Bouwmans. Subspace learning for background modeling: A survey. *Recent Patents on Computer Science*, 2(3):223–234, 2009. [Cited on page 132.]

T. Bouwmans. Traditional and recent approaches in background modeling for foreground detection: An overview. *Computer Science Review*, 11:31–66, 2014. [Cited on pages 127, 128, 130, 131, 134, and 136.]

T. Bouwmans and E. H. Zahzah. Robust pca via principal component pursuit: a review for a comparative evaluation in video surveillance. *Computer Vision and Image Understanding*, 122:22–34, 2014. [Cited on page 132.]

V. O. Z. R. Boykov, Y. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:1222–1239, 2001. [Cited on pages 84 and 301.]

P. Brault and A. Mohammad-Djafari. Bayesian segmentation and motion estimation in video sequences using a markov-potts model. *WSEAS Transactions On Mathematics*, 2004. [Cited on page 291.]

T. Brox and J. Weickert. Level set segmentation with multiple regions. *Image Processing, IEEE Transactions on*, 15(10):3213–3218, 2006. [Cited on pages 48 and 52.]

D. E. Butler, V. M. Bove Jr, and S. Sridharan. Real-time adaptive foreground/background segmentation. *EURASIP journal on applied signal processing*, 2005:2292–2304, 2005. [Cited on page 132.]

J. Canny. A computational approach to edge detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):679–698, 1986. [Cited on page 43.]

T. F. Chan, L. Vese, et al. Active contours without edges. *Image processing, IEEE transactions on*, 10 (2):266–277, 2001. [Cited on pages 42, 48, and 52.]

M. Chantler, M. Schmidt, M. Petrou, and G. McGunnigle. The effect of illuminant rotation on texture filters: Lissajous ellipses. In *Computer Vision ECCV 2002*, pages 289–303. Springer, 2002. [Cited on pages 97 and 98.]

S.-M. Chao and D.-M. Tsai. An improved anisotropic diffusion model for detail-and edge-preserving smoothing. *Pattern Recognition Letters*, 31(13):2012–2023, 2010. [Cited on page 43.]

H. Chen and B. Bhanu. 3d free-form object recognition in range images using local surface patches. *Pattern Recogn. Lett.*, 28(10):1252–1262, July 2007. [Cited on page 161.]

H.-D. Cheng, X. Jiang, Y. Sun, and J. Wang. Color image segmentation: advances and prospects. *Pattern recognition*, 34(12):2259–2281, 2001. [Cited on pages 27 and 39.]

Y. Cheng. Mean shift, mode seeking, and clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 17(8):790–799, 1995. [Cited on pages 60, 61, and 65.]

M. J. Choi, A. Torralba, and A. S. Willsky. A tree-based context model for object recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(2):240–252, 2012. [Cited on pages 251 and 261.]

C. M. Christoudias, B. Georgescu, and P. Meer. Synergism in low level vision. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 4, pages 150–155. IEEE, 2002. [Cited on pages 30, 51, and 86.]

C. Chua and R. Jarvis. Point signatures: A new representation for 3d object recognition. *International Journal of Computer Vision*, 25:63–85, 1997. [Cited on page 161.]

R. Collins. A space-sweep approach to true multi-image matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1996. [Cited on page 208.]

A. Colmenarejo, M. Escudero-Vinolo, and J. Bescos. Class-driven bayesian background modelling for video object segmentation. *Electronics letters*, 47(18):1023–1024, 2011. [Cited on page 133.]

A. Colombari and A. Fusiello. Patch-based background initialization in heavily cluttered video. *Image Processing, IEEE Transactions on*, 19(4):926–933, 2010. [Cited on page 131.]

R. V. Colque and G. Camara-Chavez. Progressive background image generation of surveillance traffic videos based on a temporal histogram ruled by a reward/penalty function. In *Graphics, Patterns and Images (Sibgrapi), 2011 24th SIBGRAPI Conference on*, pages 297–304. IEEE, 2011. [Cited on page 131.]

D. Comaniciu. An algorithm for data-driven bandwidth selection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(2):281–288, 2003. [Cited on pages 48, 52, and 66.]

D. Comaniciu and P. Meer. Mean shift analysis and applications. In *Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2*, ICCV '99, pages 1197–, Washington, DC, USA, 1999. IEEE Computer Society. ISBN 0-7695-0164-8. [Cited on pages 48, 51, 65, and 279.]

D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(5):603–619, 2002. [Cited on pages 30, 61, 63, 65, 67, and 86.]

D. Comaniciu, V. Ramesh, and P. Meer. The variable bandwidth mean shift and data-driven scale selection. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 1, pages 438–445. IEEE, 2001. [Cited on pages 37, 42, 48, 51, and 66.]

M. Cristani, M. Farenzena, D. Bloisi, and V. Murino. Background subtraction for automated multisensor surveillance: a comprehensive review. *EURASIP Journal on Advances in signal Processing*, 2010:43, 2010. [Cited on page 130.]

T. Crivelli, P. Bouthemy, B. Cernuschi-Frias, and J.-f. Yao. Simultaneous motion detection and background reconstruction with a conditional mixed-state markov random field. *International journal of computer vision*, 94(3):295–316, 2011. [Cited on page 131.]

S. A. Deadwyler and R. E. Hampson. Ensemble activity and behavior.what's the code. *Science*, 270 (5240):1316–1316, 1995. [Cited on page 164.]

P. Dollar, Z. Tu, and S. Belongie. Supervised learning of edges and object boundaries. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1964–1971. IEEE, 2006. [Cited on page 44.]

R. Dony and S. Wesolkowski. Edge detection on color images using rgb vector angles. In *Electrical and Computer Engineering, 1999 IEEE Canadian Conference on*, volume 2, pages 687–692. IEEE, 1999. [Cited on page 140.]

A. Drimbarean and P. F. Whelan. Experiments in colour texture analysis. *Pattern Recognition Letters*, 22(10):1161 – 1167, 2001. ISSN 0167-8655. [Cited on pages 43 and 102.]

R. O. Duda, P. E. Hart, et al. *Pattern classification and scene analysis*, volume 3. Wiley New York, 1973. [Cited on page 43.]

H. Eichenbaum. Thinking about brain cell assemblies. *SCIENCE-NEW YORK THEN WASHINGTON-*, 261:993–993, 1993. [Cited on page 164.]

A. Elgammal, R. Duraiswami, D. Harwood, and L. S. Davis. Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. *Proceedings of the IEEE*, 90(7): 1151–1163, 2002. [Cited on pages 132 and 295.]

A. M. Elgammal, D. Harwood, and L. S. Davis. Non-parametric model for background subtraction. In *Proceedings of the 6th European Conference on Computer Vision-Part II*, ECCV '00, pages 751–767, London, UK, 2000. Springer-Verlag. ISBN 3-540-67686-4. [Cited on page 132.]

T. Elguebaly and N. Bouguila. Bayesian learning of finite generalized gaussian mixture models on images. *Signal Processing*, 91(4):801 – 820, 2011. ISSN 0165-1684. [Cited on pages 278 and 279.]

S. Y. Elhabian, K. M. El-Sayed, and S. H. Ahmed. Moving object detection in spatial domain using background removal techniques-state-of-art. *Recent patents on computer science*, 1(1):32–54, 2008. [Cited on page 130.]

P. Elias, A. Feinstein, and C. E. Shannon. A note on the maximum flow through a network. *Information Theory, IRE Transactions on*, 2(4):117–119, 1956. [Cited on page 84.]

V. A. Epanechnikov. Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications*, 14(1):153–158, 1969. [Cited on pages 60 and 64.]

R. Evangelio and T. Sikora. Complementary background models for the detection of static and moving objects in crowded environments. In *Advanced Video and Signal-Based Surveillance (AVSS), 2011 8th IEEE International Conference on*, pages 71–76. IEEE, 2011. [Cited on page 131.]

R. H. Evangelio, M. Patzold, I. Keller, and T. Sikora. Adaptively splitted gmm with feedback improvement for the task of background subtraction. *Information Forensics and Security, IEEE Transactions on*, 9 (5):863–874, 2014. [Cited on pages 131 and 133.]

M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. Pascal 2008 results. In *2008-10-17) http://www. pascal-network. org/challenges/VOC/voc2008/workshop/index. html*, 2008. [Cited on page 56.]

S.-K. S. Fan and Y. Lin. A fast estimation method for the generalized gaussian mixture distribution on complex images. *Computer Vision and Image Understanding*, 113(7):839 – 853, 2009. ISSN 1077-3142. [Cited on pages 278 and 279.]

D. Felzenszwalb, P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59:167–181, 2004. [Cited on pages 42, 48, and 50.]

P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61:55–79, 2005. [Cited on page 161.]

P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9): 1627 –1645, sept. 2010a. [Cited on pages 160 and 161.]

P. F. Felzenszwalb, R. B. Girshick, and D. McAllester. Cascade object detection with deformable part models. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pages 2241–2248. IEEE, 2010b. [Cited on page 127.]

J. Ferryman, A. Shahrokni, et al. An overview of the pets 2009 challenge. 2009. [Cited on pages 202 and 228.]

W. Förstner and B. Moonen. A metric for covariance matrices. In *Geodesy-The Challenge of the 3rd Millennium*, pages 299–309. Springer, 2003. [Cited on page 148.]

W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *IEEE Transactions on Pattern analysis and machine intelligence*, 13(9):891–906, 1991. [Cited on page 43.]

J. Freixenet, X. Munoz, D. Raba, J. Marti, and X. Cufi. Yet another survey on image segmentation: Region and boundary information integration. In *Computer Vision-ECCV 2002*, pages 408–422. Springer, 2002. [Cited on pages 27 and 39.]

H. B. Friedrich Fraundorfer, Konrad Schindler. Piecewise planar scene reconstruction from sparse correspondences. *Image and Vision Computing*, 24:532–539, 2006. [Cited on page 203.]

A. Frome, D. Huber, R. Kolluri, T. Bulow, and J. Malik. Recognizing objects in range data using regional point descriptors. In *Proceedings of the European Conference on Computer Vision (ECCV)*, May 2004. [Cited on page 161.]

J. Fuentes-Pacheco, J. Ruiz-Ascencio, and J. M. Rendón-Mancha. Visual simultaneous localization and mapping: a survey. *Artificial Intelligence Review*, 43(1):55–81, 2015. [Cited on page 201.]

K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *Information Theory, IEEE Transactions on*, 21(1):32 – 40, Jan. 1975. ISSN 0018-9448. [Cited on pages 51, 60, 61, 63, and 279.]

J. Gallego and M. Pardàs. Region based foreground segmentation combining color and depth sensors via logarithmic opinion pool decision. *Journal of Visual Communication and Image Representation*, 25(1): 184–194, 2014. [Cited on page 183.]

D. Gallup, J.-M. Frahm, P. Mordohai, Q. Yang, and M. Pollefeys. Real-time plane-sweeping stereo with multiple sweeping directions. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, June 2007. doi: 10.1109/CVPR.2007.383245. [Cited on page 208.]

A. García and J. Bescós. Video object segmentation based on feedback schemes guided by a low-level scene ontology. In *Advanced Concepts for Intelligent Vision Systems*, pages 322–333. Springer, 2008. [Cited on pages 142, 144, 145, 151, 152, and 153.]

I. Gauthier. What constrains the organization of the ventral temporal cortex? *Trends in Cognitive Sciences*, 4(1):1–2, 2000. [Cited on page 164.]

D. M. Gavrila and V. Philomin. Real-time object detection for smart vehicles. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 1, pages 87–93. IEEE, 1999. [Cited on page 162.]

J.-M. Geusebroek, R. Van Den Boomgaard, A. W. Smeulders, and A. Dev. Color and scale: The spatial structure of color images. In *Computer Vision-ECCV 2000*, pages 331–341. Springer, 2000. [Cited on page 48.]

A. Gijsenij and T. Gevers. Color constancy using natural image statistics and scene semantics. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(4):687–698, 2011. [Cited on pages 251 and 261.]

R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 580–587. IEEE, 2014. [Cited on page 193.]

E. Goldstein. *Sensation and Perception*. Wadsworth, Belmont, CA, 2002. [Cited on pages 8, 22, 160, and 162.]

M. Gong, Y. Liang, J. Shi, W. Ma, and J. Ma. Fuzzy c-means clustering with local information and kernel metric for image segmentation. *Image Processing, IEEE Transactions on*, 22(2):573–584, 2013. [Cited on pages 32, 42, 48, and 50.]

M. A. Goodale and A. D. Milner. Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1):20–25, 1992. [Cited on page 162.]

N. Goyette, P.-M. Jodoin, F. Porikli, J. Konrad, and P. Ishwar. Changedetection. net: A new change detection benchmark dataset. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 1–8. IEEE, 2012. [Cited on page 134.]

O. G. Guleryuz. Weighted averaging for denoising with overcomplete dictionaries. *Image Processing, IEEE Transactions on*, 16(12):3020–3034, 2007. [Cited on page 112.]

R. Guo, Q. Dai, and D. Hoiem. Paired regions for shadow detection and removal. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(12):2956–2967, 2013. [Cited on pages 30 and 85.]

M. Habekost. Which color differencing equation should be used. *Int Circular Graphic Educat Res*, 6: 20–33, 2013. [Cited on pages 83 and 226.]

Y. HaCohen, E. Shechtman, D. B. Goldman, and D. Lischinski. Non-rigid dense correspondence with applications for image enhancement. *ACM Transactions on Graphics (Proceedings of ACM SIGGRAPH 2011)*, 30(4):70:1–70:9, 2011. [Cited on pages 206, 207, 211, 214, 224, 225, 226, and 229.]

M. Haindl and S. Mikeš. Texture segmentation benchmark. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE, 2008. [Cited on page 55.]

R. M. Haralock and L. G. Shapiro. *Computer and robot vision.* Addison-Wesley Longman Publishing Co., Inc., 1991. [Cited on page 23.]

R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision.* Cambridge University Press, ISBN: 0521540518, second edition, 2004. [Cited on page 211.]

M. Heikkilä and M. Pietikäinen. A texture-based method for modeling the background and detecting moving objects. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(4):657–662, 2006. [Cited on pages 96 and 134.]

S. Herrero and J. Bescós. Background subtraction techniques: systematic evaluation and comparative analysis. In *Advanced Concepts for Intelligent Vision Systems*, pages 33–42. Springer, 2009. [Cited on page 295.]

S. Hinterstoisser, C. Cagniart, S. Ilic, P. Sturm, N. Navab, P. Fua, and V. Lepetit. Gradient response maps for real-time detection of textureless objects. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(5):876–888, 2012. [Cited on pages 161, 162, 163, 170, 183, and 184.]

M. A. Hoang, J.-M. Geusebroek, and A. W. Smeulders. Color texture measurement and segmentation. *Signal processing*, 85(2):265–275, 2005. [Cited on pages 42, 47, 48, and 49.]

K. Hoffman and N. Logothetis. Cortical mechanisms of sensory learning and object recognition. *Phil. Trans. R. Soc. B*, 364:321–329, 2009. [Cited on pages 160 and 161.]

M. Hofmann, P. Tiefenbacher, and G. Rigoll. Background segmentation with feedback: The pixel-based adaptive segmenter. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 38–43. IEEE, 2012. [Cited on pages 132 and 133.]

Y. Hong, J. Yi, and D. Zhao. Improved mean shift segmentation approach for natural images. *Applied Mathematics and Computation*, 185(2):940–952, 2007. [Cited on pages 52 and 66.]

H.-H. Hsiao and J.-J. Leou. Background initialization and foreground segmentation for bootstrapping video sequences. *EURASIP Journal on Image and Video Processing*, 2013(1):12, 2013. [Cited on page 131.]

W. Hu and S.-C. Zhu. Learning a probabilistic model mixing 3d and 2d primitives for view invariant object recognition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2273 –2280, june 2010. [Cited on page 161.]

N. Hurley and S. Rickard. Comparing measures of sparsity. *Information Theory, IEEE Transactions on*, 55(10):4723–4741, Oct 2009. ISSN 0018-9448. doi: 10.1109/TIT.2009.2027527. [Cited on page 216.]

D. E. Ilea and P. F. Whelan. Ctex-an adaptive unsupervised segmentation algorithm based on color-texture coherence. *Image Processing, IEEE Transactions on*, 17(10):1926–1939, 2008. [Cited on pages 42 and 45.]

D. E. Ilea and P. F. Whelan. Image segmentation based on the integration of colour-texture descriptors-a review. *Pattern Recognition*, 44(10):2479–2501, 2011. [Cited on pages 27, 39, 46, and 49.]

Y. Ivanov, A. Bobick, and J. Liu. Fast lighting independent background subtraction. *International Journal of Computer Vision*, 37(2):199–207, 2000. [Cited on page 134.]

A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Comput. Surv.*, 31(3): 264–323, Sept. 1999. ISSN 0360-0300. doi: 10.1145/331499.331504. [Cited on page 22.]

M. Jain, H. Jegou, and P. Bouthemy. Better exploiting motion for better action recognition. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2555–2562. IEEE, 2013. [Cited on page 127.]

S. Ji and H. W. Park. Moving object segmentation in dct-based compressed video. *Electronics Letters*, 36(21):1769 –1770, oct 2000. ISSN 0013-5194. doi: 10.1049/el:20001279. [Cited on page 106.]

A. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3d scenes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 21(5):433 –449, may 1999. [Cited on page 161.]

H. Joshi and M. KhomL alSinha. A survey on image mosaicing techniques. *international Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume*, 2, 2013. [Cited on page 201.]

B. Julesz. Textons, the elements of texture perception, and their interactions. *Nature*, 290(5802):91–97, 1981. [Cited on page 96.]

M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International journal of computer vision*, 1(4):321–331, 1988. [Cited on page 52.]

E. Kasutani and A. Yamada. The mpeg-7 color layout descriptor: a compact image feature description for high-speed image/video segment retrieval. In *International Conference on Image Processing*, pages 674–677, 2001. doi: 10.1109/ICIP.2001.959135. [Cited on page 109.]

L. Kaufman and P. J. Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons, 2009. [Cited on page 66.]

F. Khan, J. Weijer, and M. Vanrell. Modulating shape features by color attention for object recognition. *International Journal of Computer Vision*, 98:49–64, 2012. [Cited on page 161.]

J.-S. Kim and K.-S. Hong. Color-texture segmentation using unsupervised graph cuts. *Pattern Recognition*, 42(5):735–750, 2009. [Cited on pages 27 and 39.]

K. Kim, T. H. Chalidabhongse, D. Harwood, and L. Davis. Real-time foreground–background segmentation using codebook model. *Real-time imaging*, 11(3):172–185, 2005. [Cited on page 132.]

W. Kim and C. Kim. Background subtraction for dynamic texture scenes using fuzzy color histograms. *Signal Processing Letters, IEEE*, 19(3):127–130, 2012. [Cited on page 132.]

T. Kinnunen, J.-K. Kamarainen, L. Lensu, and H. KäLviäInen. Unsupervised object discovery via self-organisation. *Pattern Recogn. Lett.*, 33(16):2102–2112, Dec. 2012. [Cited on page 160.]

J. J. Koenderink. The structure of images. *Biological cybernetics*, 50(5):363–370, 1984. [Cited on page 45.]

T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464 –1480, sep 1990. [Cited on page 175.]

I. Kokkinos and A. Yuille. Scale invariance without scale selection. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, June 2008. doi: 10.1109/CVPR.2008.4587798. [Cited on page 206.]

R. Krupinski and J. Purczynski. Approximated fast estimator for the shape parameter of generalized gaussian distribution. *Signal Processing*, 86(2):205 – 211, 2006. ISSN 0165-1684. [Cited on pages 278 and 279.]

S. Kudo, H. Koga, T. Yokoyama, and T. Watanabe. Robust automatic video object segmentation with graphcut assisted by surf features. In *Image Processing (ICIP), 2012 19th IEEE International Conference on*, pages 297–300. IEEE, 2012. [Cited on page 301.]

R. G. Kuehni. *Historical Development of Color Order Systems*, pages 19–103. John Wiley & Sons, Inc., 2003. ISBN 9780471432265. doi: 10.1002/0471432261.ch2. [Cited on page 26.]

K. Lai, L. Bo, X. Ren, and D. Fox. Sparse distance learning for object recognition combining rgb and depth information. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 4007 –4013, may 2011a. [Cited on page 161.]

K. Lai, L. Bo, X. Ren, and D. Fox. A large-scale hierarchical multi-view rgb-d object dataset. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 1817 –1824, may 2011b. [Cited on page 162.]

K. Lai, L. Bo, X. Ren, and D. Fox. A scalable tree-based approach for joint object and pose recognition. In *In Twenty-Fifth Conference on Artificial Intelligence (AAAI*, 2011c. [Cited on page 161.]

M. Lamarre and J. Clark. Background subtraction using competing models in the block-dct domain. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 1, pages 299 – 302 vol.1, 2002. doi: 10.1109/ICPR.2002.1044695. [Cited on pages 106 and 274.]

S. Lazebnik, C. Schmid, and J. Ponce. A sparse texture representation using local affine regions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1265–1278, 2005. [Cited on page 97.]

D. Le Gall. Mpeg: A video compression standard for multimedia applications. *Communications of the ACM*, 34(4):46–58, 1991. [Cited on page 103.]

M. Leordeanu, R. Sukthankar, and C. Sminchisescu. Efficient closed-form solution to generalized boundary detection. In *Computer Vision–ECCV 2012*, pages 516–529. Springer, 2012. [Cited on pages 44, 54, 57, and 214.]

T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *International journal of computer vision*, 43(1):29–44, 2001. [Cited on pages 97 and 98.]

L. Li, W. Huang, I. Y.-H. Gu, and Q. Tian. Statistical modeling of complex backgrounds for foreground object detection. *IEEE Transactions on image processing*, 13(11):1459–1472, 2004. [Cited on pages 134, 151, 152, 153, 278, and 285.]

Z. Li, P. Jiang, H. Ma, J. Yang, and D. Tang. A model for dynamic object segmentation with kernel density estimation based on gradient features. *Image and Vision Computing*, 27(6):817 – 823, 2009. ISSN 0262-8856. [Cited on page 96.]

Y. Liang, J. Shen, X. Dong, H. Sun, and X. Li. Video supervoxels using partially absorbing random walks. 2014. [Cited on page 153.]

T. Lindeberg. *Scale-space theory in computer vision*. Springer, 1993. [Cited on pages 45, 69, 75, 163, and 206.]

C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(5):978–994, 2011. [Cited on page 229.]

C. Lopez-Molina, M. Galar, H. Bustince, and B. D. Baets. On the impact of anisotropic diffusion on edge detection. *Pattern Recognition*, 47(1):270 – 281, 2014. ISSN 0031-3203. [Cited on page 43.]

D. G. Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999. [Cited on pages 206, 299, and 303.]

D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. [Cited on pages 163 and 166.]

N. Lyubova and D. Filliat. Developmental approach for interactive object discovery. In *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pages 1–7. IEEE, 2012. [Cited on page 183.]

J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA., 1967. [Cited on page 47.]

L. Maddalena and A. Petrosino. The sobs algorithm: what are the limits? In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 21–26. IEEE, 2012. [Cited on page 132.]

J. Malik, S. Belongie, T. Leung, and J. Shi. Contour and texture analysis for image segmentation. *International journal of computer vision*, 43(1):7–27, 2001. [Cited on pages 42, 44, 48, 54, 97, 99, and 100.]

R. Margolin, L. Zelnik-Manor, and A. Tal. How to evaluate foreground maps. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 248–255. IEEE, 2014. [Cited on pages 7 and 135.]

D. Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco, Freeman, 1982. [Cited on page 160.]

D. Marr and E. Hildreth. Theory of edge detection. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 207(1167):187–217, 1980. [Cited on page 43.]

D. Martin, C. Fowlkes, D. Tal, and J. Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 2, pages 416–423. IEEE, 2001. [Cited on pages 40, 41, 42, 54, 55, 78, 85, 112, and 119.]

D. Martin, C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(5): 530–549, 2004. ISSN 0162-8828. doi: 10.1109/TPAMI.2004.1273918. [Cited on pages 43, 44, 53, 97, 99, and 100.]

P. Meer and B. Georgescu. Edge detection with embedded confidence. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(12):1351–1365, 2001. [Cited on pages 30 and 86.]

M. Meila. Comparing clusterings: an axiomatic view. In *Proceedings of the 22nd international conference on Machine learning*, pages 577–584. ACM, 2005. [Cited on page 56.]

H. Mendez-Vazquez, E. Garcia-Reyes, and Y. Condes-Molleda. A new combination of local appearance based methods for face recognition under varying lighting conditions. In *Proceedings of the 13th Iberoamerican congress on Pattern Recognition: Progress in Pattern Recognition, Image Analysis and Applications*, CIARP '08, pages 535–542, Berlin, Heidelberg, 2008. Springer-Verlag. ISBN 978-3-540-85919-2. [Cited on page 102.]

A. Mian, M. Bennamoun, and R. Owens. On the repeatability and quality of keypoints for local feature-based 3d object retrieval from cluttered scenes. *International Journal of Computer Vision*, 89:348–361, 2010. [Cited on pages 161 and 162.]

J. Micusik, B.; Kosecka. Piecewise planar city 3d modeling from street view panoramic sequences. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. [Cited on pages 203 and 208.]

J. Micusik, B.; Kosecka. Multi-view superpixel stereo in urban environments. *International Journal of Computer Vision*, 89:106–119, 2010. [Cited on pages 206, 208, and 212.]

M. Mignotte. Segmentation by fusion of histogram-based-means clusters in different color spaces. *Image Processing, IEEE Transactions on*, 17(5):780–787, 2008. [Cited on pages 42, 48, and 49.]

K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 27(10):1615–1630, 2005. [Cited on pages 302 and 303.]

K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1):43–72, 2005. [Cited on pages 160, 166, 202, 206, and 301.]

J. Molina, M. Escudero-Viñolo, A. Signoriello, M. Pardàs, C. Ferrán, J. Bescós, F. Marqués, and J. M. Martínez. Real-time user independent hand gesture recognition from time-of-flight camera video using static and dynamic models. *Machine vision and applications*, 24(1):187–204, 2013. [Cited on page 134.]

G. Mori, S. Belongie, and J. Malik. Efficient shape matching using shape contexts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(11):1832 –1837, nov. 2005. [Cited on page 161.]

D. Mumford and J. Shah. Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on pure and applied mathematics*, 42(5):577–685, 1989. [Cited on page 42.]

S. Nadimi and B. Bhanu. Physical models for moving shadow and object detection in video. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 26(8):1079–1087, 2004. [Cited on pages 138 and 140.]

L. Nanni and A. Lumini. Coding of amino acids by texture descriptors. *Artificial Intelligence in Medicine*, 48(1):43 – 50, 2010. ISSN 0933-3657. doi: DOI:10.1016/j.artmed.2009.10.001. [Cited on page 102.]

F. Navarro, M. Escudero-Viñolo, and J. Bescos. Sp-sift: enhancing sift discrimination via super-pixel-based foreground-background segregation. *Electronics Letters*, 50:272–274(2), February 2014. ISSN 0013-5194. [Cited on pages 207, 215, 257, and 268.]

S. K. Nayar and R. M. Bolle. Reflectance based object recognition. *International Journal of Computer Vision*, 17(3):219–240, 1996. [Cited on pages 138, 139, and 142.]

A. Neri, S. Colonnese, G. Russo, and P. Talone. Automatic moving object and background separation. *Signal Processing*, 66(2):219–232, 1998. [Cited on page 127.]

W. T. Newsome, K. H. Britten, and J. A. Movshon. Neuronal correlates of a perceptual decision. *Nature*, 1989. [Cited on page 164.]

J. Nocedal and S. Wright. *Numerical optimization*. Springer Science & Business Media, 2006. [Cited on page 68.]

T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971–987, 2002. ISSN 0162-8828. [Cited on pages 96 and 282.]

M. Ozden and E. Polat. A color image segmentation approach for content-based image retrieval. *Pattern Recognition*, 40(4):1318–1325, 2007. [Cited on page 51.]

P. Paclik, R. Duin, G. van Kempen, and R. Kohlus. Supervised segmentation of textures in backscatter images. In *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, volume 2, pages 490 – 493 vol.2, 2002. doi: 10.1109/ICPR.2002.1048345. [Cited on pages 43 and 102.]

P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 12(7):629–639, 1990. [Cited on page 43.]

F. Porikli and O. Tuzel. Bayesian background modeling for foreground detection. In *Proceedings of the third ACM international workshop on Video surveillance & sensor networks*, VSSN '05, pages 55–58, New York, NY, USA, 2005. ACM. ISBN 1-59593-242-9. [Cited on pages 133, 292, 294, 295, and 298.]

H. Possegger, S. Sternig, T. Mauthner, P. M. Roth, and H. Bischof. Robust real-time tracking of multiple objects by volumetric mass densities. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. [Cited on pages 202 and 228.]

E. Potapova, M. Zillich, and M. Vincze. Learning what matters: Combining probabilistic models of 2d and 3d saliency cues. In J. Crowley, B. Draper, and M. Thonnat, editors, *Computer Vision Systems*, volume 6962 of *Lecture Notes in Computer Science*, pages 132–142. Springer Berlin / Heidelberg, 2011. [Cited on pages 183 and 185.]

A. Prati, I. Miki?, M. M. Trivedi, and R. Cucchiara. Detecting moving shadows: Formulation, algorithms and evaluation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:2003, 2003. [Cited on pages 138, 278, and 285.]

J. M. Prewitt. *Object enhancement and extraction*, volume 75. Academic Press, New York, 1970. [Cited on page 43.]

X. Qian, X.-S. Hua, P. Chen, and L. Ke. Plbp: An effective local binary patterns texture descriptor with pyramid representation. *Pattern Recognition*, 44:2502 – 2515, 2011. ISSN 0031-3203. [Cited on pages 42 and 43.]

W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971. [Cited on page 56.]

T. Randen and J. Husoy. Filtering for texture classification: a comparative study. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 21(4):291 –310, Apr. 1999. ISSN 0162-8828. [Cited on pages 42, 43, and 102.]

V. Reddy, C. Sanderson, and B. C. Lovell. A low-complexity algorithm for static background estimation from cluttered image sequences in surveillance contexts. *Journal on Image and Video Processing*, 2011: 1, 2011. [Cited on page 131.]

X. Ren. Multi-scale improves boundary detection in natural images. In *Computer Vision–ECCV 2008*, pages 533–545. Springer, 2008. [Cited on pages 44 and 98.]

L. G. Roberts. Machine perception of three-dimensional solids. Technical report, DTIC Document, 1963. [Cited on page 43.]

J. B. Roerdink and A. Meijster. The watershed transform: Definitions, algorithms and parallelization strategies. *Fundam. Inform.*, 41(1-2):187–228, 2000. [Cited on page 45.]

M. Rohrbach, S. Ebert, and B. Schiele. Transfer learning in a transductive setting. In *Advances in neural information processing systems*, pages 46–54, 2013. [Cited on page 162.]

S. Ruiz-Correa, L. Shapiro, and M. Melia. A new signature-based method for efficient 3-d object recognition. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–769 – I–776 vol.1, 2001. [Cited on page 161.]

B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1-3):157–173, 2008. [Cited on page 26.]

P. Salembier and F. Marques. Region-based representations of image and video: segmentation tools for multimedia services. *Circuits and Systems for Video Technology, IEEE Transactions on*, 9(8): 1147–1169, Dec 1999. ISSN 1051-8215. doi: 10.1109/76.809153. [Cited on pages 22, 27, and 39.]

S. Salti, F. Tombari, and L. Di Stefano. A performance evaluation of 3d keypoint detectors. In *3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2011 International Conference on*, pages 236–243. IEEE, 2011. [Cited on pages 163, 166, and 169.]

S. Salti, F. Tombari, and L. D. Stefano. Shot: Unique signatures of histograms for surface and texture description. *Computer Vision and Image Understanding*, 125:251–264, 2014. [Cited on pages 159, 162, 163, 166, 173, 183, 184, and 188.]

J. C. SanMiguel and J. M. Martínez. Shadow detection in video surveillance by maximizing agreement between independent detectors. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 1141–1144. IEEE, 2009. [Cited on page 143.]

A. Sawatzky, D. Tenbrinck, X. Jiang, and M. Burger. A variational framework for region-based segmentation incorporating physical noise models. *Journal of Mathematical Imaging and Vision*, 47(3): 179–209, 2013. [Cited on pages 42, 48, and 53.]

D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International journal of computer vision*, 47(1-3):7–42, 2002. [Cited on page 201.]

G. Schindler and F. Dellaert. Atlanta world: An expectation maximization framework for simultaneous low-level edge grouping and camera calibration in complex man-made environments. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 1, pages I–203. IEEE, 2004. [Cited on page 208.]

D. W. Scott. *Multivariate density estimation: theory, practice, and visualization.* John Wiley & Sons, 2015. [Cited on page 64.]

C. E. Shannon. *Programming a computer for playing chess.* Springer, 1988. [Cited on pages 249 and 259.]

E. Sharon, M. Galun, D. Sharon, R. Basri, and A. Brandt. Hierarchy and adaptivity in segmenting visual scenes. *Nature*, 442(7104):810–813, 2006. [Cited on pages 42, 48, 53, and 81.]

J. Shi and J. Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905, 2000. [Cited on page 54.]

L. Shi and B. Funt. Quaternion color texture segmentation. *Computer Vision and Image Understanding*, 107(1):88–96, 2007. [Cited on pages 42, 48, and 49.]

E. Simo-Serra, C. Torras, and F. M. Noguer. DaLI: Deformation and Light Invariant Descriptor. *International Journal of Computer Vision (IJCV)*, pages 1–19, 2015. [Cited on pages 206, 207, and 233.]

A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(12): 1349–1380, 2000. [Cited on page 28.]

E. S. Spelke. Principles of object perception. *Cognitive science*, 14(1):29–56, 1990. [Cited on page 163.]

P.-L. St-Charles, G.-A. Bilodeau, and R. Bergevin. Subsense: A universal change detection method with local adaptive sensitivity. *Image Processing, IEEE Transactions on*, 24(1):359–373, 2015. [Cited on pages 130, 132, 133, 134, and 135.]

C. Stauffer and W. Grimson. Adaptive background mixture models for real-time tracking. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 2:2246, 1999. ISSN 1063-6919. [Cited on pages 131, 291, and 295.]

J. Straub, G. Rosman, O. Freifeld, J. J. Leonard, and J. W. Fisher III. A Mixture of Manhattan Frames: Beyond the Manhattan World. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. [Cited on page 208.]

C. Strecha, W. von Hansen, L. Van Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. [Cited on pages 202, 223, and 228.]

A. Suga, K. Fukuda, T. Takiguchi, and Y. Ariki. Object recognition and segmentation using sift and graph cuts. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE, 2008. [Cited on page 301.]

M. Sun, H. Su, S. Savarese, and L. Fei-Fei. A multi-view probabilistic model for 3d object classes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1247–1254. IEEE, 2009. [Cited on page 164.]

Y. Tachizaki, M. Fujiyoshi, and H. Kiya. A dct coefficient sign-based background model for moving objects detection from motion jpeg coded movies. In *Intelligent Signal Processing and Communication Systems, 2009. ISPACS 2009. International Symposium on*, pages 37 –40, jan. 2009. doi: 10.1109/ISPACS.2009.5383908. [Cited on page 106.]

F. Tiburzi, M. Escudero-Vinolo, J. Bescos, and J. M. M. Sanchez. A ground truth for motion-based video-object segmentation. In *Internation Conference on Image Processing (ICIP 08)*, pages 17–20, 2008. [Cited on pages 278, 281, 282, and 295.]

E. Tola. A closed-form solution for the uniform sampling of the epipolar line via non-uniform depth sampling. Technical report, 2010. [Cited on page 221.]

E. Tola, V. Lepetit, and P. Fua. DAISY: An Efficient Dense Descriptor Applied to Wide Baseline Stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):815–830, May 2010. [Cited on pages 159, 166, 173, 188, 202, 206, 207, 211, 230, 232, 257, 268, 299, and 301.]

F. Tombari and L. Di Stefano. Object recognition in 3d scenes with occlusions and clutter by hough voting. In *Image and Video Technology (PSIVT), 2010 Fourth Pacific-Rim Symposium on*, pages 349 –355, nov. 2010. [Cited on pages 161 and 162.]

F. Tomita and S. Tsuji. *Computer analysis of visual textures*, volume 102. Springer Science & Business Media, 2013. [Cited on page 96.]

A. Torralba. Contextual priming for object detection. *International journal of computer vision*, 53(2): 169–191, 2003. [Cited on pages 251 and 261.]

K. Toyama, J. Krumm, B. Brumitt, and B. Meyers. Wallflower: Principles and practice of background maintenance. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 1, pages 255–261. IEEE, 1999. [Cited on pages 128 and 130.]

A. Treisman. *The perception of features and objects.* New York, NY, US: Clarendon Press/Oxford University Press, 1993. [Cited on pages 160 and 161.]

E. Trulls, I. Kokkinos, A. Sanfeliu, and F. Moreno-Noguer. Dense segmentation-aware descriptors. *Computer Vision and Pattern Recognition (CVPR)*, 2013. [Cited on pages 206, 207, 209, 211, 213, 214, 215, 229, 232, 257, and 268.]

D.-M. Tsai and W.-Y. Chiu. Motion detection using fourier image reconstruction. *Pattern Recognition Letters*, 29(16):2145 – 2155, 2008. ISSN 0167-8655. [Cited on page 96.]

O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. In *Computer Vision–ECCV 2006*, pages 589–600. Springer, 2006. [Cited on page 148.]

O. Tuzel, F. Porikli, and P. Meer. Pedestrian detection via classification on riemannian manifolds. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(10):1713–1727, 2008. [Cited on page 148.]

L. G. Ugarriza, E. Saber, S. R. Vantaram, V. Amuso, M. Shaw, and R. Bhaskar. Automatic image segmentation by dynamic region growth and multiresolution merging. *Image Processing, IEEE Transactions on*, 18(10):2275–2288, 2009. [Cited on pages 42, 48, 50, and 51.]

B. Ummenhofer and T. Brox. Point-based 3d reconstruction of thin objects. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 969–976, Dec 2013. doi: 10.1109/ICCV.2013.124. [Cited on page 222.]

S. R. Vantaram and E. Saber. Survey of contemporary trends in color image segmentation. *Journal of Electronic Imaging*, 21(4):040901–1, 2012. [Cited on pages 27, 39, 44, 52, and 57.]

M. K. Varanasi and B. Aazhang. Parametric generalized Gaussian density estimation. *Acoustical Society of America Journal*, 86:1404–1415, Oct. 1989. [Cited on page 279.]

M. Varma and A. Zisserman. Classifying images of materials: Achieving viewpoint and illumination independence. In *Computer Vision ECCV 2002*, pages 255–271. Springer, 2002. [Cited on page 98.]

A. Vedaldi and B. Fulkerson. Vlfeat: An open and portable library of computer vision algorithms. `http://www.vlfeat.org/`, 2008. [Cited on page 229.]

J. Vesanto. Som implementation in som toolbox. som toolbox online help, 2005. [Cited on page 176.]

J. Vesanto, J. Himberg, E. Alhoniemi, and J. Parhankangas. Self-organizing map in matlab: the som toolbox. In *In Proceedings of the Matlab DSP Conference*, pages 35–40, 2000. [Cited on page 189.]

B. Wang and P. Dudek. A fast self-tuning background subtraction algorithm. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*, pages 401–404. IEEE, 2014. [Cited on page 133.]

H. Wang and D. Suter. A novel robust statistical method for background initialization and visual surveillance. In *Computer Vision–ACCV 2006*, pages 328–337. Springer, 2006. [Cited on page 131.]

R. Wang, H. Guo, L. S. Davis, and Q. Dai. Covariance discriminative learning: A natural and efficient approach to image set classification. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2496–2503. IEEE, 2012. [Cited on page 148.]

R. Wang, F. Bunyak, G. Seetharaman, and K. Palaniappan. Static and moving object detection using flux tensor with split gaussian models. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*, pages 420–424. IEEE, 2014a. [Cited on pages 132 and 134.]

Y. Wang, P.-M. Jodoin, F. Porikli, J. Konrad, Y. Benezeth, and P. Ishwar. Cdnet 2014: An expanded change detection benchmark dataset. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*, pages 393–400. IEEE, 2014b. [Cited on pages 135 and 154.]

Z. Wang and A. C. Bovik. Mean squared error: love it or leave it? a new look at signal fidelity measures. *Signal Processing Magazine, IEEE*, 26(1):98–117, 2009. [Cited on pages 105 and 116.]

Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *Image Processing, IEEE Transactions on*, 13(4):600–612, 2004. [Cited on page 105.]

P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid. DeepFlow: Large displacement optical flow with deep matching. In *IEEE Intenational Conference on Computer Vision (ICCV)*, Sydney, Australia, Dec. 2013. [Cited on page 201.]

Y. K. Wong, J. R. Folstein, and I. Gauthier. The nature of experience determines object representations in the visual system. *Journal of Experimental Psychology: General*, 141(4):682, 2012. [Cited on page 162.]

C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: real-time tracking of the human body. In *Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on*, pages 51 –56, Oct. 1996. [Cited on page 276.]

G. Xu and Z. Zhang. *Epipolar geometry in stereo, motion and object recognition: a unified approach*, volume 6. Springer Science & Business Media, 2013. [Cited on page 201.]

Y. Xu, S. Huang, H. Ji, and C. Fermüller. Scale-space texture description on sift-like textons. *Computer Vision and Image Understanding*, 116(9):999–1013, 2012. [Cited on pages 96 and 97.]

G. Yu and J.-M. Morel. A fully affine invariant image comparison method. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 1597–1600, April 2009. doi: 10.1109/ICASSP.2009.4959904. [Cited on pages 202, 206, 207, and 229.]

Y. Yu, K. Huang, W. Chen, and T. Tan. A novel algorithm for view and illumination invariant image matching. *Image Processing, IEEE Transactions on*, 21(1):229–240, Jan 2012. ISSN 1057-7149. doi: 10.1109/TIP.2011.2160271. [Cited on pages 206 and 208.]

C. Zhang and T. Chen. A survey on image-based rendering-representation, sampling and compression. *Signal Processing: Image Communication*, 19(1):1–28, 2004. [Cited on page 201.]

H. Zhang and D. Xu. Fusing color and texture features for background model. In *Fuzzy Systems and Knowledge Discovery: Third International Conference, FSKD 2006, Xian, China, September 24-28, 2006. Proceedings*, pages 887–893. Springer, 2006. [Cited on page 134.]

H. Zhang, J. E. Fritts, and S. A. Goldman. Image segmentation evaluation: A survey of unsupervised methods. *computer vision and image understanding*, 110(2):260–280, 2008a. [Cited on pages 27 and 39.]

R. Zhang, W. Gong, A. Yaworski, and M. Greenspan. Nonparametric on-line background generation for surveillance video. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 1177–1180. IEEE, 2012. [Cited on page 131.]

S. Zhang, H. Yao, S. Liu, X. Chen, and W. Gao. A covariance-based method for dynamic background subtraction. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE, 2008b. [Cited on page 134.]

R. Zhao and W. I. Grosky. Bridging the semantic gap in image retrieval. *Distributed multimedia databases: Techniques and applications*, pages 14–36, 2002. [Cited on page 28.]

B. F. Zhenhua Wang and F. Wu. Local intensity order pattern for feature description. In *IEEE International Conference on Computer Vision (ICCV)*, pages 603–610, 2011. [Cited on page 206.]

S.-C. Zhu, C.-e. Guo, Y. Wang, and Z. Xu. What are textons? *International Journal of Computer Vision*, 62:121–143, 2005. [Cited on pages 97 and 161.]

Z. Zivkovic and F. van der Heijden. Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern recognition letters*, 27(7):773–780, 2006. [Cited on pages 132 and 133.]